Machine learning for translational medicine

Ainali, Chrysanthi

Awarding institution:
King's College London

Machine Learning For Translational Medicine

Chrysanthi Ainali

Center for Bioinformatics, Department of Informatics

King's College London, University of London

A thesis submitted for the degree of

Doctor of Philosophy

28 September 2012

To my parents and sister, for their encouragement, over all aspects,
during these years.

Acknowledgements

This thesis summarizes four adventurous and challenging years working as a Phd student in the field of Bioinformatics. During these years, I had the opportunity to collaborate with great scientists in an interdisciplinary environment and work at the borders of informatics, mathematics, medicine and biology.

First and foremost, I would like to thank my supervisor, Dr. Sophia Tsoka, for her great support and inspiration throughout every phase these years. I would also like to express my gratitude to my second supervisor, Prof. Frank Nestle, for providing expertise and insight on psoriasis and for giving me the opportunity to work very close with clinicians and biologists who helped me understand what real "translational" science is.

Special thanks go to Dr. Lazaros Papageorgiou, collaborator from UCL, for his assistance and many valuable discussions regarding the development of the HyperDM methodology. I would also like to thank Dr. Songsong Liu and Mr. Lingjian Yang for their expertise in optimisation and GAMS issues.

I am not forgetting also, Prof. Christos Ouzounis, former head of the Center for Bioinformatics, for providing his valuable feedback on part of my work and providing valuable advices.

Then, a very loud thank you to every member of the Nestle group, past and present, for their great support and help, as well the funny moments and trips we had together. Especially, I would like to thank for the collaborations we had, Dr. Gayathri Perera, Dr. Christian Hundhausen, Dr. Paola Di Meglio, Dr. Panos Karagiannis and Mr. Niwa Ali.

Of course, I cannot forget my good colleagues and friends, Ms Laura Bennett and Dr. Ignat Drozdov for useful discussions related to my work and for being there in the good times and bad times these years.

Finally, I would like to acknowledge Alexander S. Onassis Public Benefit Foundation for the financial support the last three and a half years.

.

Abstract

In biomedical sciences, the increasing amount of available high throughput data brings many challenges. The collection of such data usually results in large number of predictor genes and few samples, possibly also with high noise levels. Such problems are associated to the so-called "curse of dimensionality", i.e. the small n large p problem. Therefore, the development and application of computational protocols in bioinformatics is necessary in order to tackle these problems, translate knowledge discovery from genome-scale studies and infer new knowledge combining the different types of post-genomics data. Data mining methods, including machine learning approaches, aim to identify patterns in high-throughput data and extract information about the underlying biological interactions.

Research questions that are discussed in this thesis are disease stratification, biomarker discovery, network inference and data integration. The methodological contributions of this thesis focus on the problem encountered, nowadays, by clinicians where patients appearing to have the same disease may not respond to the same treatment. First, using supervised and unsupervised learning techniques, a machine learning strategy based on ensembles of decision trees was used to define subphenotypes based on gene expression patterns and generate potential biomarkers for disease progression. Second, we developed a network inference algorithm (NetCFS) that uses feature selection to select a number of genes (n) that are highly correlated with the phenotype of interest so as to generate n different regression problems. Third, a "top-down" approach was implemented where gene sets corresponding to biochemical pathways are used to develop a disease classification

framework. A multi-stage procedure was developed to uncover functional modules that are closely associated to the phenotype of interest and relevant to disease pathology. Phenotype-Responsive Genes (PRGs) are identified based on non-overlapping constraints of the classification procedure and association rules are used to estimate the activity level of each pathway.

Applications discussed in this thesis include skin inflammation where an integrative approach combining clinically relevant in vivo models with molecular network analysis was developed to infer disease biomarkers and to translate the rapidly growing body of data into knowledge usable at the bedside. Other disease cases studied involve cancer analyses to illustrate contributions in systems medicine. Overall, this thesis presents methodological contributions on predictive models based on machine learning techniques and mathematical programming together with relevant insights in disease mechanisms and potential treatment options.

Publications

The following list details publications related to the contributions of that thesis:

- **Ainali C.**, Nestle FO, Papageorgiou LG., Tsoka S. (May 2011) Disease Classification through Integer Optimisation. In Proceedings in 21st European Symposium in Computer Aided Process Engineering.
- **Ainali C.**, Valeyev NV, Perera GK, Williams A, Ouzounis CA, Nestle FO, Tsoka S. (2012) Transcriptome Classification Reveals Molecular Subtypes in Psoriasis. BMC Genomics,13:472, DOI: 10.1186/1471-2164-13-472
- Tsoka S., **Ainali C.**, Karagiannis P., Josephs D., Saul L., Nestle F. , Karagiannis S., (2012), Towards Prediction of Immune Mechanisms and Design of Immunotherapies in Melanoma, Critical Rev. In Biomedical Engineering
- Perera GK., **Ainali C.**, Barinaga G, Williams A, Kassen D, et al. Whole-tissue model profiling reveals IL-22 as a therapeutic target in the prevention and treatment of psoriasis with a novel role in cutaneous vascular remodelling. In Press

During the course of the Phd, I also collaborated with other scientists, which led to the following publications:

- Valeyev NV, Hundhausen C, Umezawa Y, Kotov NV, Williams G, Alex C., **Ainali C.**, Ouzounis C., Tsoka S., Nestle FO (2010) A Systems Model for Immune Cell Interactions Unravels the Mechanism of Inflammation in Human Skin. PloS Comput Biol 6(12): e1001024. Doi:10.1371/journal.pcbi.1001024.
- **Ainali C.**, Simon M., Freilich S., Espinosa O., Hazelwood L., Tsoka S., Ouzounis CA, Hancock JM. (2011) Protein coalitions in a core mammalian biochemical network linked by rapidly evolving proteins. BMC Evolutionary Biology, 11:142

- Mayr M, May D, Gordon O, Madhu B, Gilon D, Yin X, Xing Q, Drozdov I, **Ainali C**, Tsoka S, Xu Q, Griffiths J, Horrevoets A, Keshet E. (2011) Metabolic homeostasis is maintained in myocardial hibernation by adaptive changes in the transcriptome and proteome, J Mol Cell Cardiol., 50, 982-990.

- Grundberg, E, Small, KS, Hedman, AK, Nica, AC, Buil, A, Keildson, S, Bell, JT, Yang, TP, Meduri, E, Barrett, A, Nisbett, J, Sekowska, M, Wilk, A, Shin, SY, Glass, D, Travers, M, Min, JL, Ring, S, Ho, K, Thorleifsson, G, Kong, A, Thorsteindottir, U, **Ainali C**, Dimas, AS, Hassanali, N, Ingle, C, Knowles, D, et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet., 44, 10:1084-1089.

Abbreviations

AML — Acute Myeloid Leukemia
ALL — Acute Lymphoblastic Leukemia
CIO — Classification via Integer Optimization
CFS — Correlation-based Feature Selection
DDBJ — DNA Data Bank Japan
DLDA — Direct Linear Discriminant Analysis
EDTs — Ensemble of Decision Trees
GEO — Gene Expression Omnibus
GO — Gene Ontology
HyperDM — HyperBox Decision Model
KC — Keratinocyte
KEGG — Kyoto Encyclopedia of Genes and Genomes
KNN — K-Nearest Neighbor
LDA — Linear Discriminant Analysis
LP — Linear Programming
MILP — Mixed Integer Linear Programming
MPA — Mean Prediction Accuracy
MP — Mathematical Programming
MSigDB — Molecular Signature Database
NN — Normal skin tissue
PN — Non Lesional Psoriatic Sample
PP — Lesional Psoriatic Sample
PRGs — Phenotype-Responsive Genes
RF — Random Forest
RJ — Random Jungle
SRBCTs — Small Round Blue Cell Tumors
SNR — Signal-to-Noise Ratio
StD — Standard Deviation
SVMs — Support Vector Machines
SVMAttEvl — SVM Attribute Evaluator

Contents

List of Figures

List of Tables

Chapter 1

Bioinformatics and Translational Research

This chapter highlights the recent advances in genomic era that reflect on immediate advances in computational biology. Principles and scientific underpinnings of the revolution in genomic medicine are introduced and the computational challenges in biomedical research are addressed. Further, the direction of bioinformatics towards the emerging areas of integrative and translation genomics aiming at personalized medicine is outlined. Various challenges are presented and our contribution is shown.

1.1 The Genomic Era

The availability of high-throughput technology has resulted in the sequencing of the entire genome of many organisms (Lander *et al.* [2001], Venter *et al.* [2001], Zimdahl *et al.* [2004]), an emergence that has played a pivotal role in the process of biological discovery. In 2005, more than 200 complete genomes have been sequenced and made available to the research community. Powerful methods and databases have been developed to store, accumulate and analyze large-scale sequenced data. Database development was the first step for better handling and storing the available genomes. GenBank (Benson *et al.* [2012]), European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL) (Sterk *et al.*

[2007]), and DNA Data Bank of Japan ((DDBJ) (Kodama *et al.* [2012]) were the first management systems that have been developed for accumulating those data. Novel experimental techniques, such as DNA chips (Lockhart & Winzeler [2000]), serial analysis of gene expression (SAGE) (Powell [1998]) and DNA microarrays (Shyamsundar *et al.* [2005]), were initiated to monitor genome-wide mRNA levels by measuring the transcript abundance in a given sample within cells and tissues.

Currently, more than 140×10^9 nucleotides sequences and around 20,000 experiments are released in public repositories, such as DNA Data Bank of Japan (DDBJ) and ArrayExpress (**Figure 1.1A** and **Figure 1.1B**; statistics obtained from DDBJ[1] and ArrayExpress database[2], respectively). Therefore, in functional genomics research areas, data are producing by measuring the abundance of DNA and RNA molecules; thereby, researchers work on understanding the genetic basis of various evolutionary and biological processes. However, recently, advances in array design and the explosion of parallel DNA sequencing technologies facilitate the emergence of Next Generation Sequencing (NGS), which is on its way to replace microarrays by focusing on better understanding the genome function, and achieving a better quantification and annotation of trascriptomes (Shendure [2008], Wold & Myers [2008]).

In turn, the availability of whole genome sequences for various organisms and the recent advances in innovative high-throughput technologies have emerged different approaches and have changed our views on molecular biology. Whereas a few years ago, scientists were working on an individual analysis of genes or proteins, nowadays, technologies, such as genomics, transcriptomics, proteomics, metabolomics, epigenomics, have allowed the transition from the single entity analysis to a large-scale approach (Ge *et al.* [2003], Dudley & Butte [2009]). Much of the data are compressed into understandable context so as computational methods to better organize the information associated with these molecules. A move on the post-genomic era is a reality and researchers are working on achieving a complete picture of the molecular mechanistic aspects of various biological systems and complex diseases.

[1] http://www.ddbj.nig.ac.jp/

[2] http://www.ebi.ac.uk/arrayexpress/

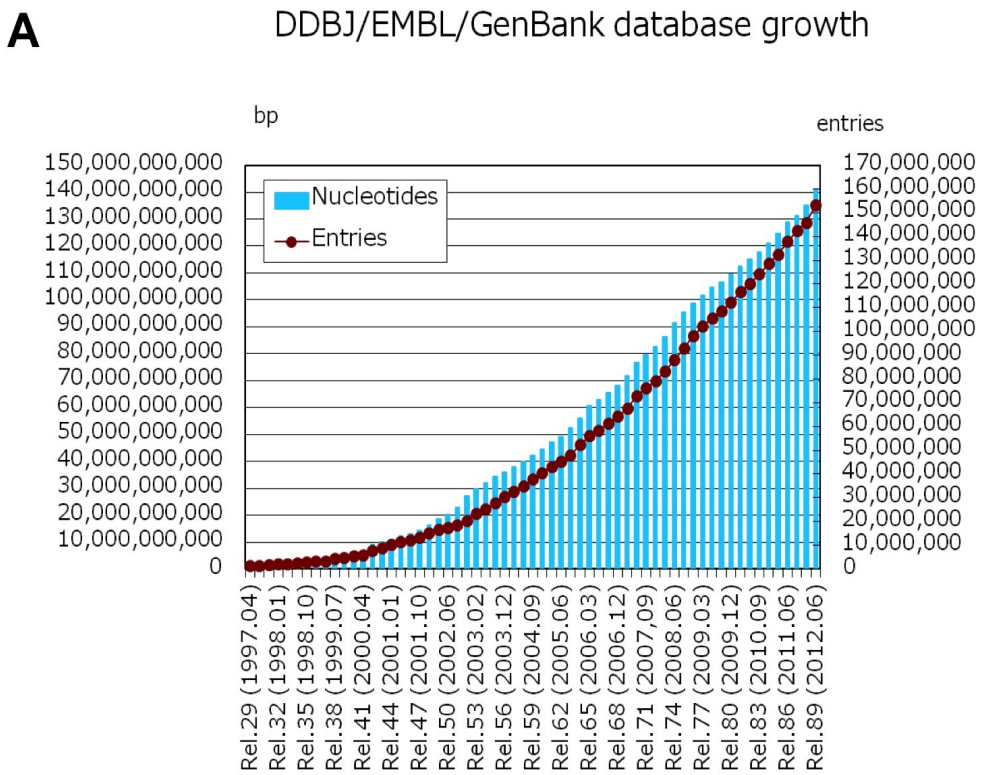

Figure 1.1: Information about public available DNA sequence and micro arrays. Graphs showing **(A)** number of nucleotide sequences that are available in DNA Data Bank of Japan (DDJB) and **(B)** the microarray experiments that are released in ArrayExpress database

1.2 Transcriptomic Data

Transcriptomics is one of the tools that are used in functional genomics to get an understanding of genes and pathways involved in complex biological systems. It is a type of large-scale experiment that through microarray technology monitors the expression levels of thousands of genes at the transcript level in parallel, under a particular condition. A microarray is a glass slide on to which DNA molecules are attached at specific locations in an order (Ziauddin & Sabatini [2001], Zhao & Bruce [2003], Simon *et al.* [2005]). These locations are called spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene (Brazma & Vilo [2000]). The DNA in the spot may either be genomic DNA (cDNA array) or a short area of oligonucleotide strands (oligo-microarray). A cDNA microarray uses the ability of a given mRNA molecule to hybridise to its original DNA coding sequence in the form of a cDNA template spotted on an array, while oligo-arrays are hybridised with only one sample and provide direct information about the expression levels in an mRNA (Brazma & Vilo [2000], Duggan *et al.* [1999], de Saizieu *et al.* [1998]).

A wide range of different microarray platforms is available today (Affymetrix[1], Agilent[2], Illumina[3]). By using this high-throughput technology, genome-wide patterns of gene expression can be studied and it allows the identification of candidate genes and their function in a given biological process. Genes showing similarity in expression pattern may be functionally related or under the same genetic control mechanism. Therefore, genome chips will pave the way for functional genomics by providing information about the expression of genes in various cell types or states as well in various diseases or treatments (e.g. how are genes regulated and how do genes and gene products interact) (Schena *et al.* [1998]).

In medical studies, microarray technology is often used as a diagnostic and prognostic tool (Mazumder & Wang [2006], Gabriele *et al.* [2006]). Large data sets from different experiments can be combined together in a single database, which allows gene expression profiles from either different samples or samples obtained using different treatments to identify additional marker genes that may

[1]http://www.affymetrix.com/estore/
[2]http://www.home.agilent.com/agilent/home.jspx?cc=GBlc=eng
[3]http://www.illumina.com/

be used to group patients into molecularly relevant categories. This increasing use of microarray technology for characterizing the transcriptional profile of disease tissue samples reveal a new tool for prediction of response to treatment and for the identification of novel therapeutic targets.

However, research in this field is challenging. Microarray data has a high dimension of variables but available datasets usually have only a small number of samples, due to the fact that is a cost-effective technology. Therefore, the task of analyzing microarray data requires novel data mining methods that are capable to handle the high dimensionality and decipher the rules in complex biological systems. Within data mining methods, machine learning is one of the main drivers of identifying the relevant pieces of information and extract knowledge. Such methods offer researchers and clinicians means to understand the underlying biology through integration of experimental data with clinical and epidemiological parameters.

Note, that recently RNA sequencing has revolutionized the exploration of gene expression. Advances in the deep-sequencing workflow, from sample preparation through data analysis, enable rapid profiling and deep investigation of the transcriptome profiling. With RNA sequencing, one can characterize all transcriptional activity, coding and non-coding, in any organism without a priori assumptions. This technique has been characterized as a revolutionary tool for transcriptomics (Wang *et al.* [2009]) and it has clear advantages in determining the structure and dynamics of the transcriptome.

1.3 Systems Biology in Translational Research

The number of sequences available as well as a wealth of experimental data in relation to disease is increasing rapidly; these data need to be understood, processed and analyzed in order to reveal the principles of how the DNA controls biological phenomena at the molecular level. Computational approaches have been developed in order to handle such enormous data influx. During the last years, research in computational genomics uses a combination of sequence analysis, algorithms and data mining approaches to predict protein function, to discover how proteins interact at the cellular level and determine the rules of proteins assembling into

different functional networks. Various methods have been developed for the analysis of datasets to find relationships (models and patterns) and to summarize the data in novel ways that are both understandable and useful to the data owner. Examples include linear equations, association rules, clusters, graphs, tree structures and recurrent patterns in time series (Yeh *et al.* [2004], Leung *et al.* [2010], Sayadi *et al.* [2011]). Methodologies have been suggested to address network inference problems, such as protein interaction networks (Tsoka & Ouzounis [2000], Jansen *et al.* [2003], Sanchez Claros & Tramontano [2012]), gene regulatory networks (Friedman *et al.* [2000]) and metabolic networks (Kanehisa [2001], Tsoka *et al.* [2004]). Methodologies have been employed to use Bayesian (Friedman *et al.* [2000], Kim *et al.* [2004]) or Petri networks (Matsuno *et al.* [2000]) or the prediction of edges between "similar" vertices (Pazos & Valencia [2001]). In fact, most of these methods base their predictions on prior knowledge, which might be based on a model of conditional dependence in the case of Bayesian networks, or the assumption that edges should connect similar vertices. The challenge is to use the noisy information that is available in an intelligent way to predict potentially new, possibly causal, relationships taking into account missing or uncertain information on genome data that hinders accurate cellular and disease modeling. Thereby, since the infancy and adolescence period of the field, bioinformatics is a fundamental core of biology and with the emergence of technological and computational developments it has established into a key discipline within every realm of life sciences and biological technology (Eisenberg *et al.* [2006], Ouzounis [2012]).

Today, with the completion of human and model organism genomes, the role of bioinformatics in post-genome era is "translational" (Butte [2008], Lussier & Li [2012]). The last five years, systems biology - a biology-based interdisciplinary study field that focuses on the systematic study of complex interactions in biological systems - play an increasing role in accelerating the translation of knowledge discovery from genome scale studies to effective detection, prognosis and treatment of complex diseases (Ideker *et al.* [2001], Kirschner [2005], Doyle *et al.* [2008], Vodovotz *et al.* [2008], Yan [2010], Fontana *et al.* [2012]). Translational research aims to aid in the development of novel techniques for the integration and transformation of clinical and biological data so as to end up in a newly

found knowledge and in solutions that could be applied in a clinical setting with potential impact on personalized medicine. These advances in personal genomics, in turn, emerge the field of translational systems biology. Translational systems biology (or translational bioinformatics) is a field that characterised of integrating clinical and biological data to enable novel knowledge by identifying molecular patterns related to a particular disease or response to treatment (Kann [2010]).

1.4 Challenges and Opportunities

It is evident that Bioinformatics plays a crucial role in translational research for biomarker discovery to prevent, detect and treat the disease (Butte [2008]). Development of algorithms and computational methods is intensively applied to biological research in order to bridge the gap between this field and the clinical applications. Different classifiers are used for the discovery of different biomarkers to stratify breast cancer patients according to the risk of having or not the disease. Diagnostic tools are being developed to detect disease based on the combination of different types of data and statistical as well network analysis is used to predict response to treatment or disease progression (Khan *et al.* [2001], Mazumder & Wang [2006], Armutlu *et al.* [2008], Dudley & Butte [2009]). However, despite that advances in computational biology and the huge data volume, the great challenge is how best to integrate this information and how to translate the insights gained from systems and network biology into clinical medicine and clinical applications so as to achieve better treatment.

1.5 Contributions

The increasing amount of genomic, proteomic, transcriptomic, metabolomic data brings many challenges, as discussed in **section 1.4**. Various types of data are available and each provides an independent and complementary view of the whole genome. Among them are gene expression, Single Nucleotide Polymorphism (SNP), copy number variation (CNV), proteomic, metabolomic and environmental and demographic data. Integration of those heterogeneous data from different sources is increasingly becoming a crucial task in understanding functions

of different components (e.g. genes, proteins), in uncovering the mechanisms of biological systems as well as in the diagnosis and prognosis of complex diseases. However, in dealing with those data is tricky. Microarray data are noisy and have missing values. The collection of those data usually results in huge number of predictor variables/genes and few samples, i.e. the small n large p problem. Therefore, the development of new approaches is necessary in order to tackle with those problems, translate knowledge discovery from genome scale studies and infer new knowledge combining the different types of post-genomic data.

The main contributions of the present thesis focus on the tasks of data integration, biomarker discovery, disease stratification and network inference. Our first contribution, which is described in **Chapter 3** concerns a procedure, based on a "top-down" approach where functional modules are used to develop a disease classification framework (Ainali *et al.* [2011]). We describe the Mixed Integer Linear Programming (MILP) mathematical model first introduced by Xu&Papageorgiou, which we further developed in order to be able to select optimal variables/genes that more effectively discriminate the different phenotypes (Hyperbox Decision Model HyperDM). Due to the limitations that arise when a gene is examined in isolation, we applied HyperDM for the first time in known functional set of genes to enhance current procedures in disease data classification and lead to improved pathway-to-disease associations. Phenotype-Responsive Genes (PRGs) are established based on non-overlapping constraints of the classification procedure and association rules are used to estimate the specificity level of association of each pathway with a phenotype of interest. Second, we developed a network inference (NetCFS) algorithm that uses feature selection to select a number of genes that are highly correlated with the phenotype of interest so as to generate different regression models (**Chapter 4**). We performed large-scale evaluation of feature selection techniques in binary and multi-class dataset and using Correlation based Feature Selection (CFS) technique, our second contribution was a method based on the identification of g correlated features to generate g sub-problems. The third methodological contribution of this thesis focus on the problem that clinicians encounter where different individuals respond different to different treatments. Clinicians require a platform that will help them to recommend drug or combination of drugs for a particular patient. In **Chapter 5**,

we propose a computational pipeline for patient stratification, using a machine learning strategy based on ensembles of decision trees (Ainali *et al.* [2012]).

Besides from the above contributions, this thesis also presents an attempt to enable the use of network biology within translational medicine proposing novel computational pipelines applied in complex skin diseases. Integrating experimental with public available data, the aim is to identify molecular signatures and putative biomarkers associated with a complex disease. We applied a protein-protein interaction approach in melanoma studies, where we integrate gene expression impaired in the experiments with public available databases to identify marker genes or gene sets that plays an underlying functional role in the effect of IgE antibody against cancer cells. Further, a co-expression approach was used to integrate the biological models of psoriasiform inflammation developed with the aim of determining whether the in vivo models were similar to human disease and to determine the effects of IL-22 on disease development and progression. This gene-gene network strategy was used to provide a better understanding of the pathogenesis of psoriasis and identify cytokine-related functional modules that could be useful for systematic identification of new promising cytokines (**Chapter 6**). A novel framework combining experimental, clinical and public data was implemented for interrogating cytokine pathways in human disease, by combining disease relevant models and integrative network analysis.

1.6 Disclaimer

Several World Wide Web sources and well known literature have been used to compile this Chapter.

Chapter 2

Fundamental Machine Learning Concepts

This chapter describes recent developments in machine learning approaches that have been applied to bioinformatics and have made significant impact on the identification of the mechanisms of various biological system components and their structural relationship. Those techniques are valuable in understanding the principles and dynamic global organization of complex biological networks.

In **section 2.1**, we focus on the world of Machine Learning (ML) by presenting advances in supervised (classification) and unsupervised learning (clustering) in the context of molecular biology and medicine. **Sections 2.2** and **2.3** present Machine Learning and Feature Selection methods used in this thesis. A brief introduction to *Support Vector Machine (SVM)*, *Ensemble Decision Trees (EDTs)* and *K-Nearest Neighbor (KNN)* algorithms as well to advances in integer optimization for classification problems are presented.

2.1 Background

The exponential growth of biomedical data available and the dynamic complexity of biological systems raise the importance of the use of computational methods. Machine Learning (ML) evolved from the field of artificial intelligence and currently the goal is to identify valid, novel, potentially useful, and ultimately

understandable hidden patterns in data (Larranaga *et al.* [2006]).

2.1.1 Supervised Learning

Supervised learning is the task to construct classifiers from a set of objects with feature vectors (input - X_1,...,X_n) and class labels (output - Y_1,...,Y_m) to learn a function $h : \{X \rightarrow Y\}$ so that $h(x)$ (hypothesis function) is a good predictor for the corresponding value of y. A Training Sample (TS) is a list of features and targets (x_k, y_k) and the training set is the set of training samples: $(x_k, y_k); \forall k = 1 \rightarrow N$. If the output is discrete-valued and we predict qualitative outputs, we call it a classification problem. If the output is continuous-valued and we predict quantitative outputs, we call it a regression problem.

Let $L : Y \times Y \rightarrow \Re$ be a loss function, that given an instance (x, y), measures the difference between the value $h(x)$ predicted by the model h from the input x, and the observed value y of the target variable. For a classification problem a typical loss function is given:

$$L(y, h(x)) = I(y \neq h(x)) \tag{2.1}$$

which is equal to 0, if $h(x) = y$ and is equal to 1 if $h(x) \neq y$.

For the regression problem, a widely used loss function is the squared error:

$$L(y, h(x)) = (y - h(x))^2 \tag{2.2}$$

The goal of supervised learning is to predict the unknown targets of some features in order to find a model h that minimizes the expected value of the loss function, taken over different instances randomly drawn from the learning samples:

$$\frac{1}{N} \sum_{k=1}^{N} L(y_k, h(x_k)) \tag{2.3}$$

To achieve this, the model should optimally fit the noise contained in the training data and lead to minimal generalization error, so as to reduce the potential of overfitting or underfitting. To estimate the generalization error of a

model either an independent dataset is used to compute the prediction error or a technique called *k-fold cross-validation* is used. In the latter case, each model is trained using the $k-1$ sets of instances and tested on the remaining sets. The generalization error is then estimated by the average prediction error over the k parts. In the case, where k is equal to the number of instances (N), then the procedure is called leave-one-out scheme (Frank *et al.* [2004], Hastie *et al.* [2009]).

Several supervised machine-learning techniques have been widely used to characterize gene expression data and select genes associated with the trait of interest (Wood *et al.* [2007]) or to classify different types of samples (Chopra *et al.* [2010]). Wood et al (Wood *et al.* [2007]) proposed a novel permutation approach to control variable selection by controlling the bias in error rates while Chopra et al (Chopra *et al.* [2010]) using nine cancer datasets and five classification algorithms, showed the improvement of the classification performance by using gene pairs instead of single genes as input. Similarly, different statistical tests were used to determine the importance of various SNPs in a disease and Support Vector Machines (SVMs) were used to to test different feature subsets and determine the one with the highest classification accuracy resulted in identifying disease associated genes and gene-gene interactions from Genome Wide Associated Studies (GWAS) (Zhou & Wang [2007]). More recently, various researchers evaluated the prediction performance of different classifiers using protein networks and pathway - based features of various datasets to predict protein structure or to examine protein-protein interactions (Chen & Liu [2005], Dutkowski & Ideker [2011], Staiger *et al.* [2012]). Support Vector Machines (SVMs), Decision Trees, Random Forest (RF), Bagging, Boosting, Nave Bayes, Logistic Regression, Artificial Neural Networks (ANN) are some of those powerful approaches that have been applied to an array of biomedical problems. Six of those were compared in order to evaluate the accuracy of the prediction of protein co-complex relationships, protein-protein interactions (PPIs) and protein co-pathway relationship, uncovering the underlying performance of the different methods in different datasets (Qi *et al.* [2006]). 30,000 yeast protein-protein pairs were selected to learn the decision model and another test set of the same size was used to evaluate the performance of the trained classifier in the context of the data set and feature encoding used. Different success rates were yielded for all classifiers with Ran-

dom Forest (RF) ranking as one of the two best methods for all combinations of the features and was further used to study the importance of different biological datasets. Statnikov and his colleagues, more recently, carried on a more comprehensive comparative benchmarking of SVM and RF algorithms (Statnikov *et al.* [2008]). They used 10-fold cross-validation scheme and applied gene selection methods in order to improve the performance. SVM exhibit good performance in most of the prediction tasks, but they suggest that, with future improvements and careful modification of the relevant classifier parameters, random forest may perform better classification. Similarly, Artificial Neural Networks (ANN) were also applied in diagnosis and detection of disease (Kapetanovic *et al.* [2004], Patel & Goyal [2007]) due to their flexibility to detect complex nonlinear relationships between dependent as well as independent variables.

2.1.2 Unsupervised Learning

In unsupervised learning, there are no explicit outputs or environmental evaluations associated with each input; given a set of inputs, the task is to discover patterns in the data above and beyond what would be considered pure unstructured noise (Hastie *et al.* [2009]). The only things that unsupervised learning methods have to work with are the observed input patterns x_i, which are often assumed to be independent samples from an underlying unknown probability distribution $P_i[x]$, and some explicit or implicit a priori information as to what is important. Thus, if we suppose that (X, Y) are random variables represented by some joint probability density $P(X, Y)$, then the goal is to directly infer the properties of this probability without any prior knowledge.

Two classic examples of unsupervised learning are clustering and dimensionality reduction. The goal of cluster analysis is to partition the observations into groups, so as to find multiple convex regions of the X-space that contain modes of $P(x)$. The goal of dimensionality reduction algorithms, such as Principal Component Analysis (PCA), multidimensional scaling, self organizing maps, is to identify low-dimensional parts within the X-space that represent high data density. This provides information about the associations among the variables and whether or not they can be considered as functions of a smaller set of la-

tent variables. Overall, unsupervised methods try to extract structure in the data, represent the data in a more compact way, or build a model of the data generating process or parts thereof.

Several methods in unsupervised learning have been used in biomedical data. Previous work include some theoretical approach of network inference by analyzing methods, such as relevance network [RELNET](Meyer *et al.* [2007], Kaleta *et al.* [2010]), CLR algorithm (Faith *et al.* [2007]) and ARACNE (Margolin *et al.* [2006]), which depends on pairwise interactions between genes, enabling the inference of large networks. Among the plethora of algorithms proposed in the literature to solve the network inference problem, there is some that can work on both time series and steady-state data, while others have been specifically designed to analyse one or the other (Bansal, Belcastro et al. 2007). Thus, the use of those techniques for network inference is envisaged to not only deal with uncertainty in input data, but to be able to handle large and complex types of information, such as complex biological systems.

In Hastie *et al.* [2009], one can find more details of supervised and unsupervised learning.

2.2 Methods in Machine Learning

This thesis is focused on predictive modeling, feature selection and network inference by using supervised and unsupervised techniques. Below, we describe the methods predominantly used in this study: Support Vector Machines (SVM), Ensemble Decision Trees (EDTs), K-Nearest Neighbors (KNN) and Mathematical Programming via Integer Optimization approaches.

2.2.1 Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised learning method from the family of kernel-based algorithms and is used for classification and regression (Risau-Gusman & Gordon [2001], Agrawal [2003]). With respect to a given collection of data, SVM calculates a maximal margin hyperplane using a particular mathematical function, namely kernel function. Kernel function maps the data in

nonlinear separable points and exploit information about the inner products between data items. The underlying idea of SVM modeling is to find the optimal line (or in N-dimensional space the hyperplane) that separates clusters of vectors in such a way that cases in one category of the target variable are on one side of the plane and cases in the other category are on the other size of the plane. The support vectors are the vectors near the hyperplane (points in the dashed lines in **Figure 2.1**). SVM classifier is robust to large number of variables, they can learn complex classification functions efficiently and they apply principles to avoid overfitting (Statnikov *et al.* [2008]). An overview of the SVM process is presented in **Figure 2.1**.

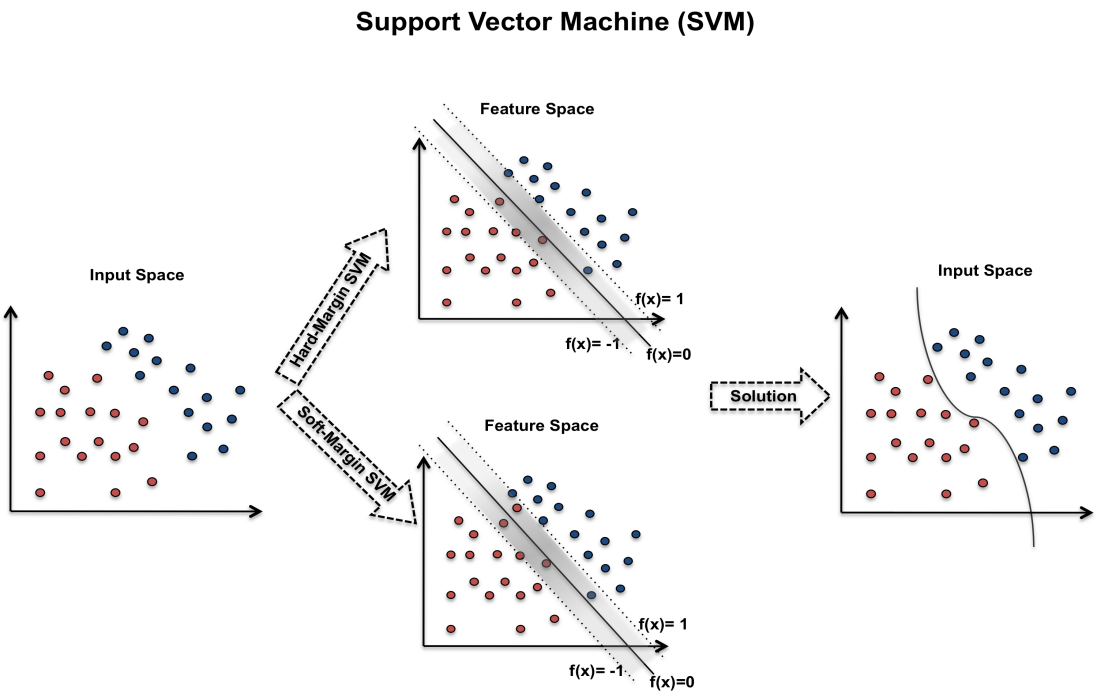

Figure 2.1: Schematic representation of a linearly separable learning sample in two dimensions. In the second graph the plain line represents the optimal hyperplane and the dashed lines are on the margin. Points located on the margin or inside the margin contribute to the predictive model.

The SVM classification function, $F(x)$, is taken the form:

$$F(x) = w \cdot x - b \tag{2.4}$$

where w is the weight vector and b is the bias computed by SVM in the training process. As shown in **Figure 2.1**, for each point x_i, the following condition should be satisfied so as to optimally separate the hyperplane and maximize the distance from the hyperplane to the closest data points (maximize the margin):

$$y_i(w \cdot x_i - b) \geq 1, \forall(x_i, y_i) \in D \tag{2.5}$$

When x_i are the closest vectors, $F(x)$ is equal to 1 (as in **Figure 2.1**) and the margin is equal to $\frac{1}{||w||}$. For hard-margin SVM (**Figure 2.1**), support vectors are the points, which are "on the margin". Therefore, the training problem becomes a constrained optimization problem aiming at minimizing $||w||$.

Minimize:

$$Q(w) = \frac{1}{2}||w||^2 \tag{2.6}$$

subject to:

$$y_i(w \cdot x_i - b) \geq 1, \forall(x_i, y_i) \in D \tag{2.7}$$

However, not all the classification problems are linearly separable. In such cases, soft-margin SVM (**Figure 2.1**) introduces a slack variable ξ_i, which estimates and minimizes the degree of misclassification while maximizing the margin. **C** is a parameter that determines the tradeoff between the margin size and the amount of error in training.

Minimize:

$$Q(w, b, \xi_i) = \frac{1}{2}||w||^2 + C\sum_i \xi_i \tag{2.8}$$

subject to:

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i, \forall(x_i, y_i) \in D, \xi_i \geq 0 \tag{2.9}$$

During the last years, Support Vector Machines (SVMs) have been widely used in the field of computational biology due to the global and unique solution that can provide comparing for example with Neural Networks (NN) that suffer from multiple local minima. In 1999, SVM classifier was applied to predicting a

protein's structural class from its amino acid sequence (Jaakkola *et al.* [1999]). A new approach was developed (SVM−Fisher) that used a discriminative framework built on top of a generic model (e.g. HMM) in order to identify remote homologies. From the accuracy point of view, it performed better than any other previous method. More recently, another study determines how to classify genes according to the similarity on their switching mechanisms using the Fisher kernel (Pavlidis *et al.* [2001]). The same year, Hua and Sun use SVM to perform prediction of subcellular localization according to protein classes (Hua & Sun [2001]). One year later, Zavaljevski and Reifman describe the application of SVM in order to distinguish benign and pathologic human immunoglobulin light chains (Zavaljevski *et al.* [2002]). SVM has also been used to solve the difficult multi-class problem of classifying a sequence of amino acids into one of many known protein structural classes (Melvin *et al.* [2007]). At the same time, a considerable attention has been given to the analysis of microarray gene expression. In 2008, linear classifiers were used to separate the data (splice sites) with hyperplanes and SVMs as well as other related kernel methods approved that provide high accuracy and can deal with high dimensional and large datasets (Ben-Hur *et al.* [2008]). Lately, the algorithm have been used via a network based approach to predict clinical outcome of patients and resulted in significantly improved reproducibility of prediction performance across different data sets (Chen *et al.* [2011]).

2.2.2 Ensemble of Decision Trees (EDTs)

Tree-based methods, such as classification and regression trees (CART) (Breiman *et al.* [1984]), partition the feature space into a set of rectangles and then fit a tree structured model in each one. An example is shown in **Figure 2.2**. Decision tree is grown using the training set and the process is binary; at each level, the parent node is splitting into two daughter nodes in a recursive way. The algorithm chooses the locally best discriminatory feature at each stage in the process. Given a feature vector, the algorithm is looking for a feature that best separates the data minimizing the mean impurity of the two partitions. A measurement of impurity of each partitioning is made and the query that generates the least impure partitions is selected. This process is applied recursively on each sub-

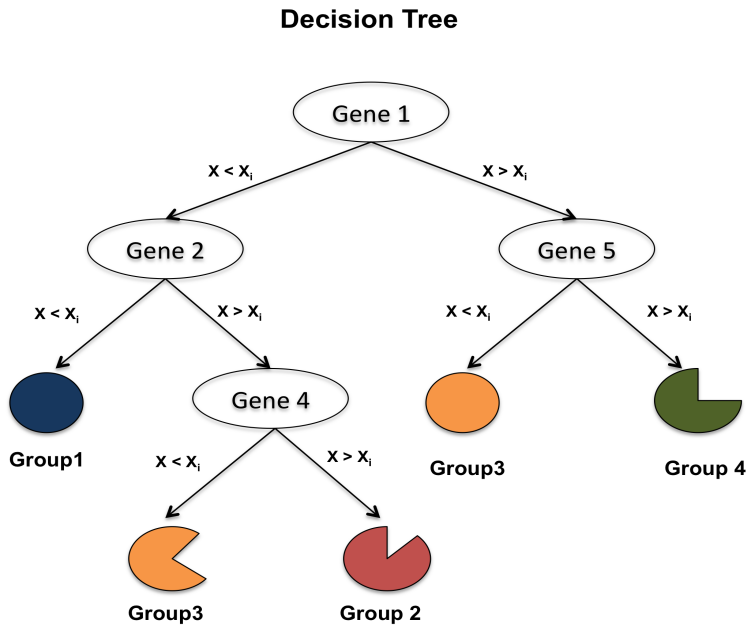

Figure 2.2: Schematic representation of a decision tree using gene expression data. Each internal node of the tree (ellipse) represents the input variables which are used to recursively split the learning sample resulting in the terminal nodes (colored circles), that contain the output variable. Internal nodes represent the 5 genes (input variables) and the terminal nodes correspond to the predicted group (blue circle for group 1, red for group 2, orange for group 3 and green for group 4). Full and semi-circles represent the percentage of predicted group.

partition until either a homogeneous (perfect classification) or non-homogeneous terminal nodes (the ends of the tree) will be achieved. Overall, the basic idea of a tree classifier is to reduce as much as possible the uncertainty about the output variable in the resulting subsets of samples, so as to capture complex interaction structures in the data with low bias.

In machine learning, a committee of base classifiers has been shown to improve overall prediction accuracy and can form a superior classifier compared to individual ones (Tan & Gilbert [2003], Wang [2006]). Bagging, Boosting and Random Forest (RF) are examples of such ensemble techniques, where a population of de-

cision trees is first developed from the training data each with a different feature subset, and then they are combined to form the composite predictor (**Figure 2.3**). In this way, these methods can achieve more accurate classification with the increase of model complexity as well as offer higher stability and better generalization of unknown data. Therefore, ensemble algorithms, such as Ensemble of Decision Trees (EDTs), impose regularization for effective analysis in "large p, small n" problems which are widely found in many bioinformatics applications.

As shown in **Figure 2.3**, Ensemble of Decision Trees (EDTs) are grown non-deterministically using a two-stage randomization procedure to build a large collection of de-correlated trees and then averages them to achieve low variance. By default, decision trees are suitable for the perturbation procedure applied to the training data and as base classifiers are diverse from each other (each classifier makes a misclassification independently)(Tsymbal A [2005]) . Bagging (**Figure 2.3a**) and Random Forests (**Figure 2.3c**) use perturbed data sets and different feature sets for training tree classifiers while in boosting (**Figure 2.3b**) each classifier is trained and combined from the samples with different classification weights. Therefore, the greedy nature of one-step-at-a-time node splitting enables trees (and hence forests) to obtain a different classification hypothesis and the "grouping property" of trees enables to deal with correlation and interaction among variables.

In comparison to bagging and boosting, Random Forests (RF) has been shown to be more effective in addressing the gap between the sample size and data dimension generated by high-throughput biological experiments (Diaz-Uriarte & Alvarez de Andres [2006]). In this thesis, Random Forest (RF) is applied in microarray data classification and development of random forests variants was also implemented. The construction of random forests is described with the pseudocode in **Appendix A**. Assuming that the training set is consisting of N observations, then N cases are sampled with replacement from the original dataset randomly (this is called "bagging") so as, some observations will be selected more than once, and others will not be selected. Therefore, by this selection, when the training set for the current tree is drawn by sampling, about one-third of the cases are left out of the sample. This set of out-of-bag data (OOB) is used to get a running estimate of the classification error as trees are added to the forest. After

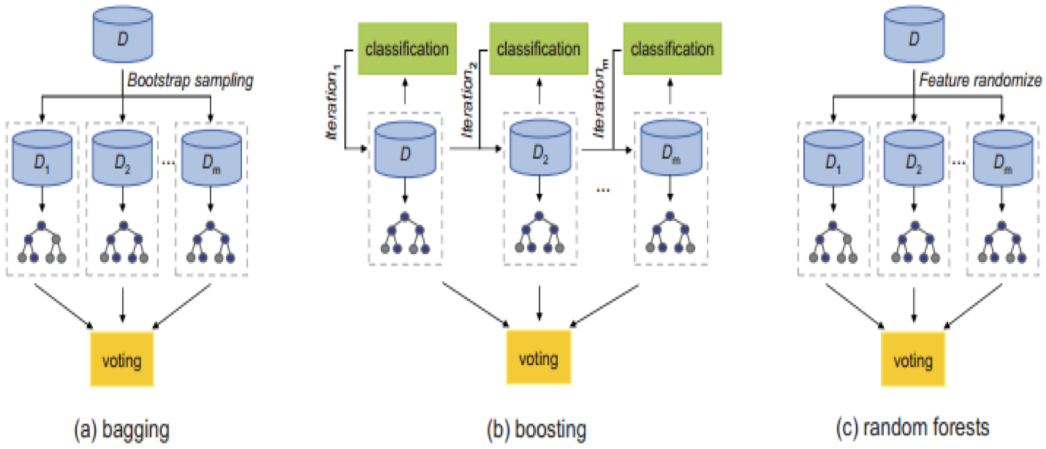

Figure 2.3: Schematic representation of EDTs methods. (a) Bagging: bootstrap sampling is applied to the training set (D) and different feature sets (D_1, D_2, ... D_m) are used for training tree classifiers , (b) Boosting: each classifier is trained separately and is combined from the samples with different classification weights. (c) Random Forests: features are sampled by replacement and new features sets are used to train different decision trees in parallel and their predictions are combined to make the overall prediction by voting for the most popular class. (Figure adapted from Yang [2010])

the construction of each tree using a different bootstrap sample from the original data, only a subset of the total set of predictor variables is considered as possible splitters for each node. The best predictors, though, will not be considered for each split, but a predictor excluded from one split may be used for another split in the same tree. In other words, a large number of decision trees are grown in parallel, and their predictions are combined to make the overall prediction for the forest by voting for the most popular class.

An important feature of Random Forest (RF) is the proximity matrix. Random Forest (RF) proximity is determined by examining the terminal node membership of the data. If sample i and sample j both fall within the same terminal node of a given tree, the proximity between i and j is increased by one. Summing over all terminal nodes in a forest produces the proximity matrix, which represents the degree of similarity between sample points. After transforming the RF proximity matrix to a dissimilarity matrix, it opens the door to many clustering and visualization approaches for detecting data structures. Thus, these properties of RF make it an appropriate tool for genomic data analysis and bioinformatics research.

These unique advantages offered by ensemble methods resulted in their wide application to many bioinformatics problems in dealing with small sample size, high-dimensionality, and complexity data structure. Bagging and boosting methods have been initially applied to classify tumors using gene expression proles (Ben-Hur *et al.* [2008], Dudoit & Fridlyand [2002]). Similarly, Tan and Gilbert used seven publicly available datasets and proved that ensemble methods of bagging and boosting are more robust and accurate in microarray data classification comparing to single tree classifier (Tan & Gilbert [2003]). More recently, nonlinear and non-parametric Random Forest approach has been successfully applied to various problems, e.g. genetic epidemiology and microbiology, medical diagnosis (Yang *et al.* [2009]), pathway clustering (Pang & Zhao [2008]) protein-protein interactions and mapping human metabolic pathways (Macchiarulo *et al.* [2009]). Pang et al. develop a Random Forests-based approach to identify how informative genes within pathways are related to each other from gene expression data and investigate possible links between pathways that do not have a shared gene (Pang & Zhao [2008]). Shi et al. successfully used RF unsupervised learning for

tumor class discovery based on immunohistochemical tumor maker expression (Shi *et al.* [2005]). Therefore, ensemble based classification is promising for the microarray analysis.

2.2.3 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is one of the most straightforward instance-based learning algorithms that require no model to be fit. Given a query instance x_0, the algorithm locates the k training points x_n, $n = 1, \ldots, k$, closest in distance to x_0, and then classify using majority vote among the k neighbors (Geva & Sitte [1991], Hastie *et al.* [2009]). Training points (instances) are mapped within an n-dimensional feature space where each of the n-dimensions corresponds to one of the n-variables. The similarity between instances is measured using a distance metric, such as Euclidean distance. Therefore, the objective of k-nearest neighbors is to minimize the distance between two similarly classified instances, while maximizing the distance between instances of different classes. Note that choosing a more suitable distance metric for a specific dataset can improve the accuracy of instance-based classifiers, although their major disadvantage is the large computational time for classification.

Nearest neighbor classifiers has been applied in a number of real domains, such as bioinformatics, and has the ability to perform well on data sets that are not linearly separable, often giving better performance than more complex methods in many applications (Dudoit & Fridlyand [2002]). A comparison of KNN with other classification methods has shown that the algorithm produce robust performance results similarly with the others when is used for the classification of tumor samples using gene expression profiles (Li *et al.* [2001]). Graph based KNN is previously used in protein interaction prediction as well in protein structure prediction (Shen & Chou [2006]).

2.2.4 Classification via Integer Optimization (CIO)

Mathematical programming (MP) makes use of recent advances in optimization theory and has been used in applications for the efficient solution of the classification problems. Comparing with other supervised learning approaches, MP

methods are straightforward to implement through standard modeling tools, do not make strict assumptions about the distribution of the data analyzed and no parameter optimization is required during the training process. Linear programming formulations are widely used to extract patterns hidden from gene expression data to achieve tissue classification or disease gene identification and biomarker discovery (Niijima & Okuno [2009], Daruwala *et al.* [2004]). Mixed Integer Programming classifiers (MIP) extend the classification model by adding binary variables indicating misclassified data points and solving datasets involving a relatively small number of training samples. Glen (Glen [2001]) presented a Mixed Integer Linear Programming (MILP) model to maximise the classification accuracy for two-group classification problems. Later, piecewise linear classifiers were applied to approximate nonlinear discriminant functions to improve the classification performance (Glen [2003], Ryoo [2006]). Finally, in 2006, Uney and Turkay introduced an MILP model hyper-box classifiers which has been applied in protein fold type prediction (Y. *et al.* [2008]), drug classification (Dagliyan *et al.* [2009], Armutlu *et al.* [2008]) and tumor classification (Dagliyan *et al.* [2011]). More recently, Xu and Papageorgiou (Xu & Papageorgiou [2009]), develop a completely different MILP formulation, which involves significantly fewer binary/continuous optimisation variables and constraints and can identify disjoint regions of the dataset through hyper-boxes with the best prediction accuracy. However, while there has been extensive studies undertaken in mathematical programming (MP) approaches for discriminant and classification analysis, they focus mainly on two-class classification techniques. Therefore, there is a need of robust classification methods with generalization to multiple-class problems.

Formulating biomarker identication problems as MP models and finding effective methods to solve these problems are important steps in these classification studies. Comparing with Linear Programming (LP) models and statistical approaches, Mixed integer-programming classifiers (MIP)-based methods may obtain better performances, while the existence of binary variables make MIP models solve datasets involving a relatively small number of training samples. The optimisation strategy differentiates samples that belong to multiple classes in feature space through hyper-boxes. The objective is to minimise the misclassified samples and performance of the model is improved through an iterative

solution. The formulation of the MILP is detailed below:

Minimize:

$$\sum_{s}(1 - E_s) \tag{2.10}$$

Subject to:

Hyperbox Enclosing Constraints

$$A_{sm} \geq X_{im} - \frac{LE_{im}}{2} - U(1 - E_s), \forall s, i, m \tag{2.11}$$

$$A_{sm} \leq X_{im} + \frac{LE_{im}}{2} + U(1 - E_s), \forall s, i, m \tag{2.12}$$

Non Overlapping Constraints

$$X_{im} - X_{jm} + U_{ijm} \geq \frac{LE_{im} + LE_{jm}}{2} + \varepsilon \tag{2.13}$$

$$\sum_{m=1}^{M}(Y_{ijm} + Y_{jim}) \leq 2M - 1, \forall i = 1, \ldots, N - 1, j = 1 + i, \ldots, N \tag{2.14}$$

$E, Y_{ijm} \in 0, 1; LE_{im \geq 0}; X_{im} : unrestricted$

where for boxes i and j on attribute m, Y_{ijm} is a binary variable demonstrating whether box i and j overlap. X_{im} is the central coordinate of hyper-box i on attribute m. LE_{im} is the length of hyper-box i on attribute m. E_s indicates whether sample s is correctly classified. ε is defined as the small positive number that prevents the overlap of two boxes. Parameter A_{sm} represents the value of sample s on attribute m. U is a suitable upper bound.

For the improvement of the model accuracy, an iterative solution procedure is suggested and multiple boxes for each phenotype were generated. A testing scheme is further performed where new samples are assigned to existing hyper-boxes so as to identify their class memberships according to pathway signatures. Thus a disease sample is allocated to a class according to the distance between sample s and box i on attribute m, $DIST_{s,i,m}$:

$$DIST_{s,i,m} = max(0, A_{sm} - UB_{im}, LB_{im} - A_{sm}), \forall s, i, m \tag{2.15}$$

where UB_{im} and LB_{im} are the upper and lower bound of the box i on attribute m, respectively. They are used for extracting the association learning rules in order to define the phenotype margins according to the expression gene measurements within the pathway markers.

$$LB_{im} = X_{im} - \frac{LE_{im}}{2}, \forall i, m \tag{2.16}$$

$$UB_{im} = X_{im} + \frac{LE_{im}}{2}, \forall i, m \tag{2.17}$$

The distance between testing sample s and hyper-box i, DSI_{si}, is defined as:

$$DSI_{si} = \sqrt{\sum_{m=1}^{M} DIST_{sim}^2} \tag{2.18}$$

An extension and an application of the model of Xu and Papageorgiou (**?**) is presented in **Chapter 3** using a hyper-box representation (Hyperbox Decision Model - HyperDM).

2.3 Feature Selection Methods

Large number of features and a limited number of observations in some datasets, such as microarray data, are not useful for producing a desired learning result and may lead the learning algorithm to overfit to the noise. Reducing the number of features (dimensionality) is important in statistical learning; it can save storage, computation time and increase comprehensibility (Ma & Huang [2008]). Therefore, the implementation of feature selection methods is important and attempts to solve this curse of dimensionality problem by selecting through different ways a subset of features that are useful to build a good predictor (Ben-Dor *et al.* [2000], Golub *et al.* [1999], Pan [2002]).

The objective of feature selection methods is to achieve a subset of features (genes that are good diagnostic indicators; biomarkers), which eliminates noise

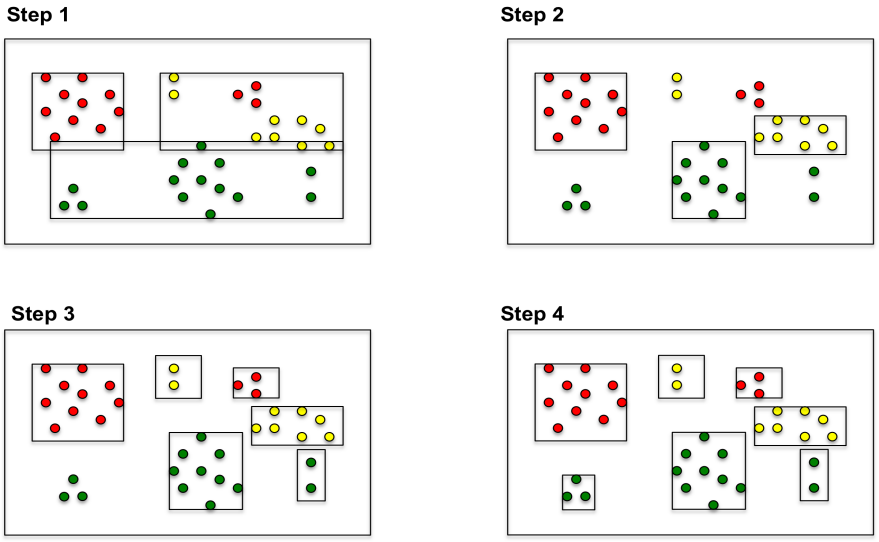

Figure 2.4: Schematic representation of the steps of HyperDM method in a two-dimensional three-classes example. Each color represents one class. **Step 1** determines the boundaries for corresponding classes for all samples. In **Step 2**, the MILP model is applied to the training data, hyper-boxes are constructed and samples outside of the boxes are identified. Then, one more box for each class with misclassified samples is added (**Step 3**). Finally, the algorithm follows an iterative procedure by solving single level MILPs until the objective function values of two successive iterations are the same (**Step 4**)

from the classification task, optimises classification between different classes, avoids overfitting and provides faster and more cost-effective predictors. Filters, wrappers, and feature weighting are the three main approaches in feature (gene) selection. Filter methods remove features that are irrelevant with the training data according to some prior knowledge. Wrapper approaches are dependent to classification algorithms to evaluate the feature subsets and as a result, they have high computational complexity. Feature weighting methods simply rank features according to a relevant weigh. As shown in **Chapter 3**, the following feature scoring functions are further emploeyd to find a small set of features (markers) that best explains the difference between different phenotypes/conditions: *Correlation based-Feature Selection (CFS), Significance Attribute Evaluator, Support Vector Machine Attribute Evaluator (SVMAttrEval)* and *Attribute Ranking* from Weka [1] machine learning package (Frank *et al.* [2004]), and *Variable Importance Metrics (VIM)* derived from Ensemble of Decision Trees (EDTs) classifiers.

2.3.1 Correlation based-Feature Selection (CFS)

Correlation-based feature selection (CFS) algorithm was first introduced as a filter method by Hall & Smith [1998]. The algorithm is based on the hypothesis that *'a good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other'* (Hall & Smith [1998]). In other words, a feature is useful if it is correlated with the class; otherwise it is irrelevant. If the correlation between each of the features and the class is known, and the inter-correlation between each pair of features is given, then the CFS evaluates a subset of features by estimating:

$$CFS_s = \frac{k\overline{r}_{cf}}{\sqrt{k + k(k-1)\overline{r}_{ff}}} \tag{2.19}$$

where CFS_s is the score of a feature subset S containing k features, $\overline{r}_{cf}$ is the average of the correlations between the features and the class ($f \in S$), and $\overline{r}_{ff}$ is the average inter-correlation between the features. The equation measures the redundancy among the group of features and indicates the predictive ability of a

[1] http://www.cs.waikato.ac.nz/ml/weka/

subset of features (Hall & Smith [1998]).

2.3.2 SVM Attribute Evaluator (SVMAttributeEval)

SVM Attribute Evaluator algorithm is a wrapper feature selection method that evaluates the importance of a feature by using an SVM classifier. As discussed in **Section 2.2.1**, a property of SVM is that the weights w_i of the classification function $F(x)$ are a metric of a small subset of the training examples that are closest to the decision boundary and lie on the margin, called 'support vectors'. Therefore, in SVMAttributeEval the features are ranked by the square of the weight assigned to them by the SVM (Guyon *et al.* [2002]). The algorithm is following an iterative procedure, which ranks features and discards the worst one, after sub sampling of the dataset similarly like the Recursive Feature Elimination (RFE) algorithm:

1. An SVM is created on the given data sets
2. The objective function is applied for all features (this is used as the ranking criteria).
3. The feature with the lowest value is removed and added to the feature list
4. The SVM is recreated with the remaining features
5. The above steps are repeated for all the features
6. A list of ranked features is returned.

More details can be found in Guyon *et al.* [2002].

2.3.3 Variable Importance Measures (VIMs) in Ensemble of Decision Trees (EDTs)

Ensemble of Decision Trees (EDTs) can be used to select subset of features by estimating the importance of each predictor variable. That serves as means to quantify the effect of each variable (i.e. gene) in discriminating among the relevant class outcomes (phenotypes). The Random Forest (RF) approach serves two fundamentally different Variable Importance Measures (VIMs): the Gini importance and the permutation importance (Diaz-Uriarte & Alvarez de Andres

[2006]). Mean decrease in accuracy is motivated from statistical permutation tests and is based on the decrease of prediction accuracy when values of a variable in a node of a tree are permuted randomly (Strobl *et al.* [2008]). The Gini Index measure (GI) is derived from training of the RF classifier and is used as the splitting criterion in the construction of the trees (Menze *et al.* [2009]). The Gini Index VIM is a simple average across all the trees in the forest of the decrease in node impurity observed using that predictor.

Naturally, a gene is likely to be diagnostic if its expression value in a disease state is different from each expression value in a normal state (when the gene predictive power to distinguish between different classes is high). Consequently, there is a need to eliminate the redundant genes more effectively, to choose and to rank them based upon their ability to distinguish between various classes of samples. If one considers the terms "discriminative" and "inequality" to have analogous meaning, the selection of a gene against a patient population can be evaluated in statistical terms by the Gini coefficient (Graczyk [2007]).

2.4 Disclaimer

Well known literature has been used to compile this Chapter. As well, the book of Hastie *et al.* [2009], was used for detailed explanation of supervised and unsupervised learning.

Chapter 3

Disease Classification

This chapter presents the state-of-the art of disease classification using microarray gene expression data. Human disease classification dates to the late 19th century, but recently, there is a growing number of transcriptomic, metabolomic, proteomic and genomics data identifying molecular underpinnings of many disorders. Therefore, integrating all these data for probing biological systems can guide new biological insights.

In **section 3.1**, we present the results of prediction performance of various classification methods. In **section 3.2**, we introduce a variant mathematical approach of *HyperDM* (HyperBox Decision Model) for disease classification through integer optimization. *HyperDM* classifier was first applied in a gene expression-based approach and then in a pathway-based approach.

3.1 Disease Classification using Microarray Gene Expression Data

Cellular expression patterns obtained through microarray technologies are crucial in defining the mechanisms underlying different disease phenotypes (disease states, outcomes or responses). Molecular signatures or markers are extracted from genome-wide expression profiles by marking how well gene(s) can differentiate between these disease phenotypes or between disease samples and appropriate controls. In this context, classification of gene expression has been widely used to

discover diagnostic markers and dissect molecular signatures of complex diseases (Golub *et al.* [1999], van 't Veer *et al.* [2003], Liu *et al.* [2005], Hwang *et al.* [2008] Cun & Frohlich [2012]), that determine the disease state of samples or predict response to therapy.

3.1.1 Introduction

One of the main challenges in computational biology is the extraction of useful information from biological data. As described in **Chapter 2**, **section 2.2**, in a classification problem, a set of data points is divided into classes. Given an instance of the set, the classifier is used to assign labels to new instances according to the value of its predictor variables and a set of decision rules. Different classification models have been successfully applied in disease classification on the basis of molecular profiles. In particular, supervised machine learning has been used to great effect in numerous bioinformatics prediction methods and related applications continue to grow in biomedical literature (Jensen & Bateman [2011]).

Expression microarrays have been widely used when studying classification algorithms; to classify biological samples in a number of novel ways such as by tumor type (Golub *et al.* [1999]), toxicological mode of action (Thomas *et al.* [2001]), and pharmacological mechanism (Gunther *et al.* [2003]). For example, Golub and his colleagues constructed the weighted voting classifier to distinguish between two types of human acute leukemia's: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) (Golub *et al.* [1999]). From a total of 6817 genes obtained from Affymetrix GeneChip, 50 genes were selected on the basis of Signal to Noise Ratio (SNR) in 38 training samples and a voting algorithm was applied to classify 34 new samples in the testing data set as either AML or ALL. Tibshirani *et al.* [2002] used the nearest shrunken centroid method for the multi-class prediction of the small round blue cell tumors (SRBCTs) data, where a set of 43 genes achieved comparable classification performance. Dudoit et al. compared the performance of various predictors including Linear Discriminant Analysis (LDA), Classification And Regression Trees (CART) and K-Nearest Neighbors (KNN) using three DNA microarray data sets: a leukemia (ALL/AML) data (Golub

et al. [1999]), a lymphoma data (Alizadeh *et al.* [2011]) and the 60 cancer cell line (NCI-60 panel) data (Ross *et al.* [2000]). They concluded that most of the classifiers perform similarly across data sets, with a remarkable performance of diagonal Linear Discriminant Analysis (LDA) and K-Nearest Neighbors (KNN) compared to more sophisticated methods such as aggregated classification trees. Diaz-Uriarte & Alvarez de Andres [2006] investigated the use of Random Forest (RF) for classification of microarray data and showed that an ensemble of decision trees forest has comparable performance in multi-class data to other classification methods, including Direct Linear Discriminant Analysis (DLDA), K-Nearest Neighbors (KNN) and Support Vector Machine (SVM).

Therefore, a variety of statistical and machine learning methods have been used to analyze the high dimensional data generated by high-throughput technologies for disease classification and biomarker discovery. These methods have been shown to have clinical relevance in cancer detection for a variety of tumour types. The purpose of this study is first to gain insight on the performance of various classification methods, facilitate the comparison of different methods and to investigate the effect of classification techniques on the prediction of disease subtypes based on gene expression datasets.

3.1.2 Real and Synthetic Datasets

The classifiers discussed in **Chapter 2** were applied using both model - based synthetic data and real data with different number of classes, attributes (genes) and observations (samples) (Table 3.1). Raw data were normalized using quantile normalization and expression estimates were computed using the Robust Multi-chip Average (RMA) method (Irizarry *et al.* [2003]). For each dataset, the first goal was to assess the performance of classification methods. The datasets used are the following:

- Leukaemia Dataset (Golub *et al.* [1999]): The dataset contains 3050 genes subdivided in two classes: 27 samples with acute lymphoblastic leukaemia (ALL) and 11 samples with Acute Myeloid Leukaemia (AML).

- Small Round Blue Cell Tumors (SRBCTs) (Khan *et al.* [2001]): The data

consists of expression measurements on 6,567 genes (2,308 genes after filtering for minimal level of expression) obtained from glass-slide cDNA microarrays. The tumors are classified as non-Hodgkin lymphoma (NHL), Ewing sarcoma (EWS), neuroblastoma (NB), or rhabdomyosarcoma (RMS).

- Synthetic Dataset (Yeung & Bumgarner [2003]): The dataset consists of four-classes, with each class represented by 20 samples and each sample is described by 1000 gene expression profiles. This four-class classification problem was generated by allocating a set of genes as patterned genes for class description. Non-patterned genes from real dataset are then added. Noise is subsequently added to represent the microarray experiments. Noise is generated under two constraints: biological and technical. One of the challenges the researchers faced was finding the best ratios for noise variables. To deal with this issue, the researchers created a number of datasets with different variable values.

Table 3.1: Binary, multi-class and synthetic datasets used for the evaluation of the classification methods

Name	Type	No Genes	No Samples	No Classes
Binary Dataset	Leukemia	3051	38	2
Multi-class Dataset	Cancer	2308	64	4
Synthetic Dataset	Synthetic	1000	80	4

3.1.3 Evaluation and Comparison of Classification Methods

A classification procedure is performed within the datasets described in **Section 3.1.2** to illustrate the potential of various methods to elucidate the concerted actions of different genes in disease. To assess predictive performance, we trained the following classifiers: Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbour (KNN), Bagging, Boosting and HyperDM. A description of each methodology is provided in **Chapter 2**. The machine learning algorithms,

such as Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbour (KNN), Bagging and Boosting, were implemented in WEKA, by LIBSVM, Random Forest, IBK, Bagging and AdaBoost functions, respectively. The mixed integer linear programming model (HyperDM) and the iterative solution are implemented in the General Algebric Modeling System (GAMS) on a CentOS 5.2 64 bit UNIX environment using the CPLEX mixed integer optimisation solver with 1% margin of optimality.

For the different classifier parameters, we implemented classification methods described above and obtained the prediction performances shown in **Figure 3.1**. To assess predictive performance, each classifier is trained using 10-fold cross validation. This scenario was repeated 50 times and the mean prediction accuracy is reported. Support Vector Machine (SVM) was trained using linear, polynomial, radial and sigmoid kernel functions. Random Forest (RF) was trained building ensemble of decision trees (EDTs) with small parameter values (depth, features and trees) and subsequently generating a number of more complex forests. Increasing these parameters improves the prediction but not in all the datasets. K-Nearest Neighbour (KNN) was implemented with different K values. For the real datasets the best results were obtained where K had the smallest value ($K = 1$). Because it is more difficult for the classifier to separate instances when the number of features is large, increasing the K value may reduce class prediction. Bagging and Boosting were used as wrappers to other classifiers, including Random Forest and Decision Stamp under default settings. Default settings (using random forest and decision tree) produce good results for synthetic and binary datasets and a higher prediction rate is observed of random forest under bagging. Interestingly, this result is comparatively better than what is achieved using more complex random forest classifiers. This indicates that selecting features with replacement increases the efficiency of the algorithm. From the experiments conducted we can also observe how the bag size effects the results obtained. Overall, Support Vector Machine (SVM) outperform in binary datasets while Random Forest (RF) perform well in multi-class dataset.

As shown in **Figure 3.1**, the well-known classifiers usually require parameter optimization to obtain accurate results depending on the structure of the data. In contrast, as described in **Chapter 2**, **section 2.2.4** , HyperDM approach

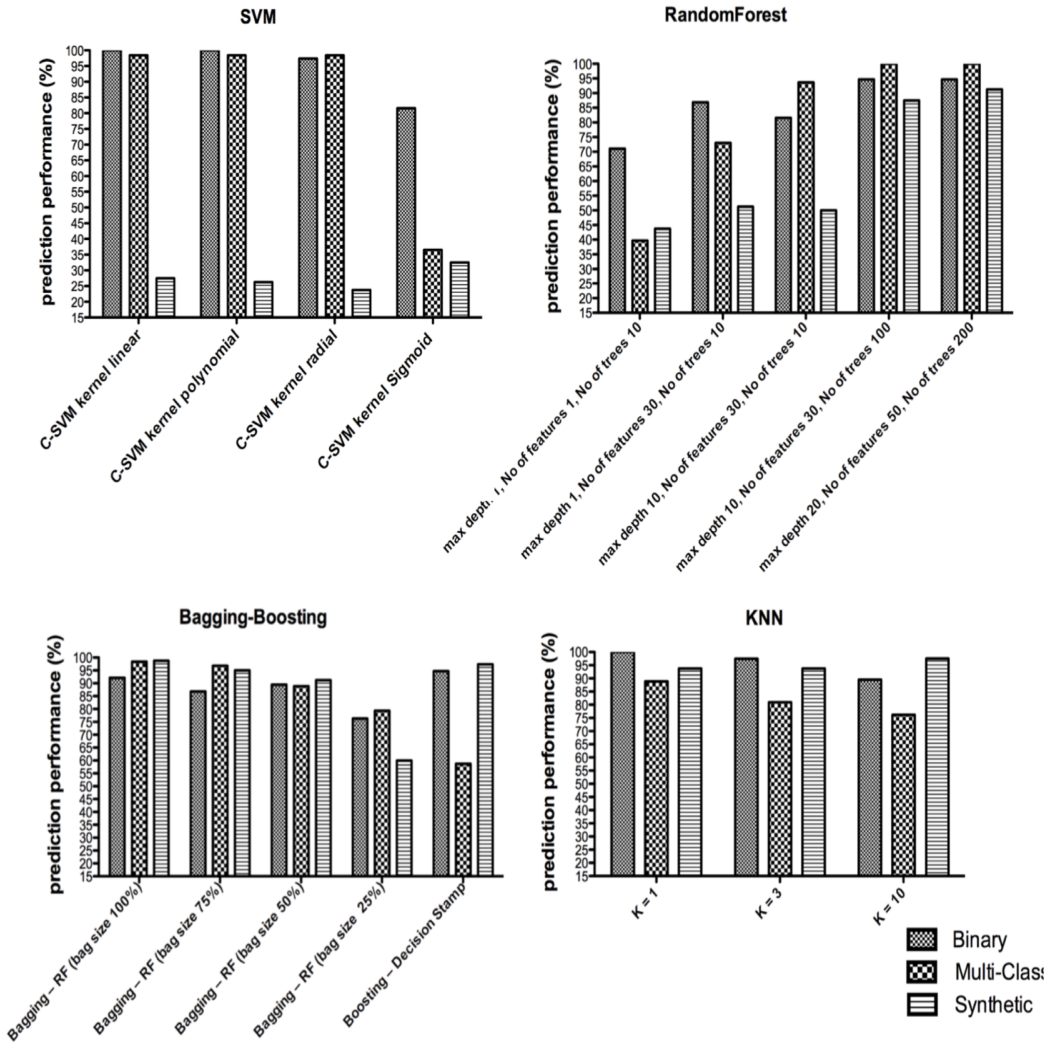

Figure 3.1: Predictive performance of (A) SVM, (B) Random Forest, (C) bagging and boosting and (D) KNN algorithms

does not require parameters to be optimized in order to obtain high classification accuracies and can be used for different types of data without any modifications. However, there is a limitation when using a dataset with large number of variables (as in the case of microarrays) and small number of samples. The inclusion of redundant information may weaken performance of some machine learning algorithms. In that case, the algorithm terminates due to resource limit, may overfit the data and cannot achieve the optimal solution (Wang *et al.* [2005b], Chuang *et al.* [2011]). In **section 3.1.4**, a hybrid approach that integrates feature selection with classification methods is introduced. This strategy could achieve consistent prediction accuracy across different datasets (Li *et al.* [2004]) and assist in a better understanding of disease mechanisms (Guyon *et al.* [2002]).

3.1.4 Hybrid Classification Approach through Feature Selection

While predictive models could be based on the expression of more than a few tens of genes, several reasons motivate the search for short lists of predictive genes (biomarkers). As discussed previously, selecting a small subset of genes for classification can improve the prediction performance, provide a reduced number of marker genes that could be used for diagnostic purposes and assist in disease prognosis (Wang *et al.* [2005b]). An important characteristic of microarray data is the large number of genes relative to the number of samples. However, this high dimensionality in gene space increases the computational complexity while it usually decreases the accuracy of the classification, since usually a small number of the genes are correlated with the phenotype of interest. Thus, a prediction model built from thousands of available predictor variables and a relatively small sample size can be quite unstable (Miller [2002]). This fact brings the necessity of gene selection or gene reduction for the high dimensional gene space.

Here is proposed a robust and effective hybrid classification pipeline based on feature selection, extraction of new predictors list and training of classifiers with the new dataset. From a machine learning perspective, restricting the number of variables is often a way to reduce over-fitting and can thus lead to unbiased predictions on new samples. From a biological viewpoint, inspecting the genes

selected in the signature may shed light on biological processes involved in the disease, understand the molecular mechanisms of the disease and suggest novel targets. A schematic diagram of the hybrid approach is shown in **Figure 3.2**. Using as input dataset the whole gene expression matrix ($g \times s$ matrix, where g is the number of genes and s is the number of samples), first are applied different feature selection methods to extract predictors that better discriminate the samples and may be more informative for the phenotypes of interest. Then, a new expression profile matrix is generated using only the genes subset identified in the previous step and different classifiers are trained on the new dataset. Finally, evaluation of the classifiers is performed and if the prediction accuracy is lower than 70%, then feature selection functions are re-applied on the expression matrix and the new selected feature subsets are used to generate a new matrix to train the classifier. The choice of a threshold of 70% prediction performance was applied due to the fact that increasing the error rate, the signal to noise ratio of the model is decreasing too, indicating more noise than signal in the experiment. It has been shown through our experiments that performance greater than 70% results in in a classification model with higher signal to noise ratio.

New gene expression matrices are generated using different feature selection algorithms found in WEKA machine learning package (Frank *et al.* [2004]) as well Variable Importance Metrics (VIM) generated by Random Jungle, the computational and memory efficient implementation of RF (Schwarz *et al.* [2010]). For efficiency purposes, we also generated random datasets, selecting 10, 25 and 50 features, respectively, to investigate the feature selection techniques. For the generation of the random dataset, 100 permutations of samples were implemented in Java.

In Weka, four feature selection methods were examined: (i) Correlation based Feature Selection (CFS) (ii) Significance Attribute Evaluator, which evaluates the importance of a feature by computing the Probabilistic Significance. (iii) Support Vector Machine Attribute Evaluator (SVMAttrEval) and (iv) Attribute Ranking, which rank the features based on how effective they are when we construct a random forest. More details are described in **Chapter 2**, **section 2.3**. However, different feature selection methods result in different gene ranking and as well in different classification rules. Therefore it is reasonable to combine the information

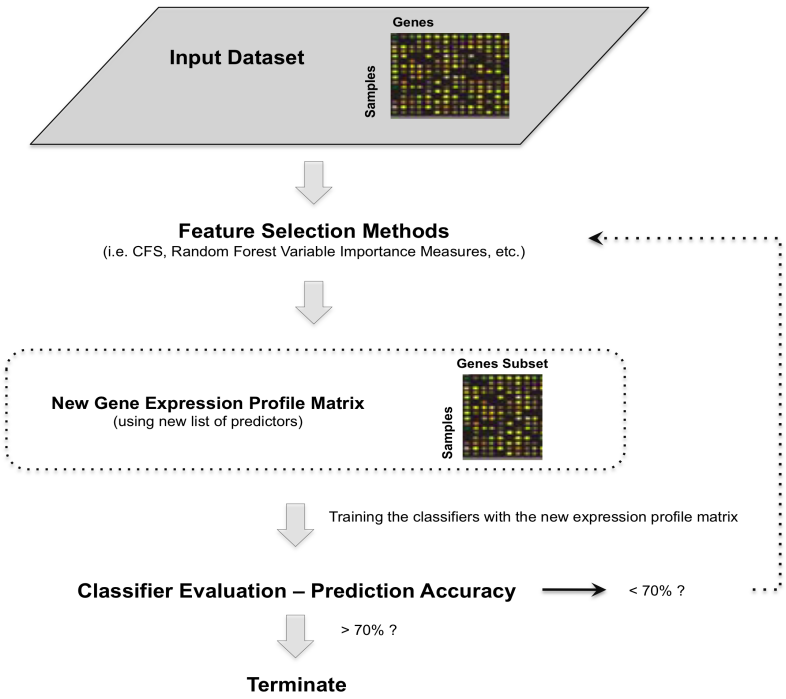

Figure 3.2: Schematic diagram of the hybrid classification approach based on feature selection.

from several methods by aggregating the selections of a number of feature scoring functions. Thus, a consensus ranking was computed according to a Markov Chain rank aggregation method as applied elsewhere (Dutkowski & Gambin [2007]) This method consists of gathering the top ranked genes within a specified threshold across all datasets. The threshold of 100 was applied in all of the datasets used.

Note, that using the ranked list derived from SVM Attribute Selection method and the Random Jungle Variable Importance, four models were constructed and solved by selecting 10, 25, 50 and 100 features (biomarkers) respectively for further processing.

Once the feature sets are completed, they are used to create classifiers with different parameters. We compare and contrast the results of the original dataset with these new reduced datasets in order to check in which cases machine learning algorithms have an advantage over simpler procedures. Tables 3.2,3.3 and 3.4 show the accuracies achieved by the standard model (original dataset) and the hybrid model for the binary, multi-class and synthetic dataset, respectively. In all the tables, the classification accuracy is evaluated using SVM (sigmoid and radial kernel functions), RF, KNN, Bagging, Boosting and HyperDM for the original dataset (without feature selection), as well for the datasets derived after feature selection. The experimental results show that the accuracy of microarray data improves when variable selection is implemented comparing those without feature selection.

A very interesting observation is that HyperDM perform well in different types of data and the performance is comparable to other methods described above (SVM, RF, KNN, Bagging and Boosting). In contrast, the other classification methods, even with the application of the hybrid model, do not perform similarly well on all the datasets. HyperDM model exhibits advantages, in terms of handling disjoint sets of samples well and performing feature prioritisation through extraction of rules that may increase prediction performance.

3.1.5 Discussion

The large dimensionality (up to several tens of thousands of genes) and the small sample sizes, in addition to experimental complications like noise and variability

Table 3.2: Prediction accuracies in binary dataset

Dataset	SVM Sigmoid	SVM Radial	RF small selection	RF automatic selection	KNN $K = 1$	KNN $K = 10$	Bagging	Boosting	HyperDM
Original Dataset[a]	81.57	97.37	71.05	89.47	100	89.5	92.1	94.73	72.41
Combined all -8	97.3	100	97.3	97.37	97.3	100	97.36	97.36	99.60
Combined - 22	100	100	94.73	94.74	100	100	92.1	97.36	100
Random - 50	79.63	89.00	72.36	84.73	88.47	84.00	87.15	86.94	65.54
Random - 25	76.00	85.21	72.73	82.11	83.31	82.89	86.10	83.84	60.32
Random - 10	74.31	78.00	72.21	79.89	75.47	76.21	80.68	78.73	65.43
CFS - 48	100	100	97.36	100	100	100	100	94.73	99.58
SVMAttEvl 10	100	100	97.36	97.36	100	100	97.36	97.36	99.58
SVMAttEvl - 25	100	100	92.10	94.73	92.10	100	97.36	97.36	97.5
SVMAttEvl - 50	100	100	94.73	94.73	100	100	100	97.36	94.23
SVMAttEvl-100	100	100	92.10	100	100	100	97.36	100	85.23
RJ 10	100	100	97.36	100	100	97.36	100	97.36	98.4
RJ 25	100	100	100	100	100	100	100	97.36	94.23
RJ 50	100	100	94.73	97.36	100	100	97.36	97.36	92.33
RJ- 100	100	100	94.73	94.73	100	100	100	97.36	82.10

[a]Classification models were run without applying feature selection

Table 3.3: Prediction accuracies in multi-class dataset

Dataset	SVM Sigmoid	SVM Radial	RF small selection	RF automatic selection	KNN $K = 1$	KNN $K = 10$	Bagging	Boosting	HyperDM
Original Dataset[a]	81.57	98.41	39.68	85.71	88.89	76.19	98.41	58.73	66.8
Combined all -11	61.90	93.65	73.01	92.06	95.23	96.82	92.06	58.73	95.32
Combined - 35	88.88	96.82	74.60	98.41	100	100	100	58.73	99.22
Random - 50	40.73	75.69	42.13	73.54	71.34	72.42	83.46	50.17	54.3
Random - 25	40.34	63.98	43.23	68.47	69.39	65.47	75.26	48.96	53.32
Random - 10	40.07	54.82	40.30	60.40	57.7	56.06	63.03	57.76	48.7
CFS - 48	93.65	95.23	95.23	100	100	100	100	58.73	98.68
SVMAttEvl 10	95.23	100	80.95	93.65	100	100	100	63.49	100
SVMAttEvl - 25	100	100	79.36	95.23	100	100	100	63.49	98.43
SVMAttEvl - 50	85.73	100	74.60	100	100	100	100	61.90	90.27
SVMAttEvl-100	100	98.41	71.42	100	100	100	100	61.90	87.23
RJ 10	58.73	92.06	68.25	88.88	96.82	85.71	79.36	60.31	98.67
RJ 25	69.84	96.82	74.60	96.82	100	98.41	100	60.31	98.54
RJ 50	84.12	96.82	74.60	98.41	100	100	100	60.31	90.27
RJ- 100	90.47	95.23	69.84	96.82	100	100	100	60.31	86.3

[a]Classification models were run without applying feature selection

Table 3.4: Prediction accuracies in synthetic dataset

Dataset	SVM Sigmoid	SVM Radial	RF small selection	RF automatic selection	KNN $K = 1$	KNN $K = 10$	Bagging	Boosting	HyperDM
Original Dataset[a]	27.5	23.75	43.75	28.75	93.75	97.5	74	31.25	26.17
CFS	65	55	30	55	62.5	57.5	47.5	22.6	65.75
Random - 50	36.45	32.3	27.35	20.9	30.8	30.45	31.9	28.75	35.14
Random - 25	38.25	42.15	29.75	38.25	33	35.35	37	31.3	32.4
Random - 10	37.5	46.15	29.3	36.5	36.5	41.4	39.55	31.45	31.98
SVMAttEvl 10	50	35	30	37.5	37.5	50	60	25	88.3
SVMAttEvl - 25	62.5	60	37.5	42.5	45	52.5	60	25	82.32
SVMAttEvl - 50	70	67.5	42.5	47.5	60	67.5	70	25	80.4
RJ 10	75	72.5	66.25	72.5	72.5	78.75	75	36.25	90.32
RJ 25	88.75	90	61.25	67.5	85	97.5	82.5	31.25	88.24
RJ 50	88.75	90	56.25	76.25	90	97.5	93.75	31.25	87.96
RJ- 100	90	93.75	50	63.75	91.35	97.5	95.5	31.25	90.21

[a]Classification models were run without applying feature selection

render the analysis of microarray data a new challenge of research in bioinformatics (Somorjai *et al.* [2003]). The work described in this part of the thesis is related to the impact of high dimensionality in supervised learning. Classification methods using all the features do not necessarily perform well due to the existence of many irrelevant noisy features that do not contribute to the reduction of classification error. Feature selection techniques, though, preserve the original semantics of the variables and select a subset of them that are relevant (Wang *et al.* [2005c]). Therefore, a hybrid model that integrates feature reduction methods with evaluation algorithms can result in high predictive performance.

The experiments on the microarray datasets highlighted that the classification accuracy achieved with different strategies is highly sensitive to the type of data and to the precise values of the learning parameters of the algorithms. For example, increasing the number of trees and the number of nodes in the Random Forests method induces higher prediction performance (**Figure 3.1)**. Thus it is of great interest to find a way to automatically tune the parameters of the classification algorithms, to search for the best feature set or to develop an approach that will not require parameters to optimize in order to obtain high predictive accuracies.

An effective hybrid classification approach is implemented to investigate the relationship between various feature reduction methods and the resulting classification performance. The hybrid model, introduced in the previous section, is applied on publicly available data sets including leukemia, SRBCT and one synthetic dataset. The results underline the importance of HyperDM, which perform well without any parameter modification and capture the variability of the different data. In addition, this study emphasizes that various feature selection techniques return common features varying only in their relevance score. Therefore, a feature selection approach is proposed to enhance the robustness of the selected predictor genes. Based on the observations of Tables 3.2, 3.3 and 3.4, highest accuracy in classification is established when using the consensus features comparing with the accuracy in the case of using the full feature set. This effort might provide more biologically meaningful predictors with less false positives and more accurate diagnostics. The strong influence of different feature reduction methods on the classification accuracy observed highlights the need for

further investigation in the complex interaction between feature reduction and classification.

In conclusion, the experimental results show that by employing the proposed hybrid model fewer gene subsets needed to be selected and better classification accuracy could be obtained. However, expression-based classification can be challenging in complex diseases due to factors such as cellular and genetic heterogeneity. Therefore, a more effective way to increase the discriminatory power of a model may be to combine gene expression measurements over groups of genes that fall within functional modules, such as pathways (Draghici *et al.* [2003], Subramanian *et al.* [2005], Tian *et al.* [2005], Chuang *et al.* [2007]). The adaptation of these procedures to solve this problem would be an interesting direction of future research and is discussed in the following section, **section 3.2)**

3.2 Pathway-based Disease Classification through Integer Optimisation

As discussed in the previous section, among classification methods used for the analysis of disease data are bayesian predictors, support vector machines, decision trees and mathematical programming (Furey *et al.* [2000],Statnikov *et al.* [2008], Armutlu *et al.* [2008]). The majority of such approaches are based either on considering all genes simultaneously or a single gene at a time. However, effects such as genetic dissimilarities across patients and cellular heterogeneity within tissues are important hurdles in obtaining good predictive performance (Ein-Dor *et al.* [2005], Lee & Tzou [2009], Symmans *et al.* [1995]), resulting in weak associations among genes and decreased classification performance (Saeys *et al.* [2007]). The concept of using sets of genes reflecting specific biochemical pathways has been proposed as a promising alternative over both integrative (all-gene) or reductionist (single-gene) analyses, so as to account for the joint effects of genes involved in similar functional groups and, therefore, derive stronger pathway-to-phenotype associations (Schadt [2009]).

Stemming from the above, a key challenge in disease classification studies is the development of robust pipelines to relate genotypic information to disease

phenotypes through known molecular interactions. Common resources to record biochemical interactions have been used in this context, such as Gene Ontology (GO) (Ashburner *et al.* [2000]), KEGG (kyoto encyclopedia of genes and genomes) (Kanehisa *et al.* [2004]), Reactome (Croft *et al.* [2011]) and BioCarta [1]. Research activities to investigate the implication of such gene sets in disease based on transcriptomics include the gene set enrichment analysis (GSEA) (Subramanian *et al.* [2005]), the use of pathway-based random forest classification (Pang & Zhao [2008]) and hierarchical clustering (Gatza *et al.* [2010]). These methods use all member genes within each pathway and then establish the predictive power of those modules to the disease classification problem. However, such approaches depend heavily on pathway definitions, which may not entirely correspond to the specific disease under study. It is, therefore, an open question for the classification procedure to detect the gene sets that best reflect which diagnostic markers are relevant to the phenotype of interest. In this respect, methods to determine which gene sets correspond to particular disease class have been proposed recently and are based on classification-driven pathway activity measures and combinatorial boolean rules (Chuang *et al.* [2007]Lee *et al.* [2008], Vaske *et al.* [2010], Park *et al.* [2010]).

A novel pipeline for disease data classification through a pathway-based analysis is reported. The Mixed Integer Optimisation Model (HyperDM) is extended to classify tissue samples (breast cancer and posriasis samples) to appropriate disease classes by separating feature space using hyper-box principles (Xu & Papageorgiou [2009]). Advantages of the proposed classification model pipeline are (i) good descriptive power through the extraction of classification rules and (ii) effectiveness in modeling disjoint data regions. The method presented here implements a feature selection process and the use of appropriate rules to characterize gene sets in disease classification is also reported. Overall, this work aims to enhance current procedures in disease data classification through pathway gene-sets and lead to improved pathway or genes-to-phenotype associations.

[1] http://www.biocarta.com/

3.2.1 Dataset

Breast cancer is the most frequently diagnosed malignancy and has been intensively studied by gene expression profiling (Cheang *et al.* [2008]). Psoriasis is a systemic, inflammatory autoimmune skin disease affecting 2-3% of the world population (Nestle *et al.* [2009]). Both are complex diseases with unknown molecular mechanisms and interesting

Breast tumor tissues of 721 patients were used from NCBI Gene Expression Omnibus (Edgar *et al.* [2002]) (GEO[1]; GSE1561, GSE2034, GSE3494, GSE1456) and were tested on Affymetrix platforms U133A and U133B. Each sample was recorded as a vector of 22,215 probeset expression values, which were then mapped to the corresponding genes for further interpretation. For psoriasis tissue samples, two datasets were considered: the first dataset consisted of 21 biopsies from healthy donors (NN) and 28 paired non-lesional (PN) and lesional plaque type psoriatic patients (PP) (Yao *et al.* [2008]), and the second dataset comprised of 64 NN, and 58 PN and PP (Gudjonsson *et al.* [2010]). In the first dataset, there are also 5 samples from psoriasis patients that only provided lesional skin biopsies. Both datasets were acquired using HGU133plus2 Affymetrix chips. According to the original publications, each dataset was divided into two or three populations of distinct phenotypes defining binary and multi-class problems (Table 3.5).

Pathway-specific gene expression profiles were built from microarray measurements using pathway-gene associations from KEGG database (Kanehisa *et al.* [2004]) and data from the C2 functional set were used and downloaded from the Molecular Signature Database (MSigDB database v3.0, September 2010). From this set, the model is applied on 186 pathways that covered 5267 genes, where not all of them could be found on the datasets, due to the various platforms used.

3.2.2 HyperBox Decision Model (HyperDM)

A multi-stage procedure is applied to integrate expression and pathway datasets for phenotype prediction of breast cancer samples according to pathway signatures (**Figure 3.3**). Expression profiles for m tumor samples obtained from breast cancer patients across k probes are summarized by a $m \times k$ matrix, where

[1] http://www.ncbi.nlm.nih.gov/geo

Table 3.5: Four independent breast cancer and two psoriasis datasets. Two binary and four-multiclass

Dataset	GEO Ref.	Samples	Class	Samples per Class
Farmer *et al.* [2005]	GSE1561	49	3	6 Apocrine
				27 Luminal
				16 Basal
Wang *et al.* [2005a]	GSE2034	286	2	209 ER+
				77 ER-
Miller *et al.* [2005]	GSE3494	247	2	213 ER+
				34 ER-
Pawitan *et al.* [2005]	GSE1456	124	4	37 Normal like
				25 Basal
				15 ERBB2
				62 Luminal (A, B)
Yao *et al.* [2008]	GSE14905	77	3	21 Normal (NN)
				28 peri-lesional (PN)
				28 lesional(PP)
Gudjonsson *et al.* [2010]	GSE13355	180	3	4 Normal (NN)
				58 peri-lesional (PN)
				58 lesional(PP)

each entry (S_{mk}) denotes the expression level of gene k in patient m. Using gene-to-pathway correspondence, pathway-specific tumor expression matrices are created, where samples, features and classes are used as input to the classification algorithm.

For the multi-class disease classification problem, Mixed Integer Linear Programming (MILP) model was used, employing a hyper-box representation, HyperDM (Xu & Papageorgiou [2009]). The model is described in **Chapter 2**. The proposed mathematical model and the iterative solution was implemented in the General Algebric Modeling System (GAMS) using the CPLEX mixed integer optimization solver [1] with 1% margin of optimality. All experiments were run

[1]The CPLEX Optimizer was named for the simplex method as implemented in the C programming language, although today it also supports other types of mathematical optimization and offers interfaces other than just C (http://en.wikipedia.org/wiki/CPLEX).

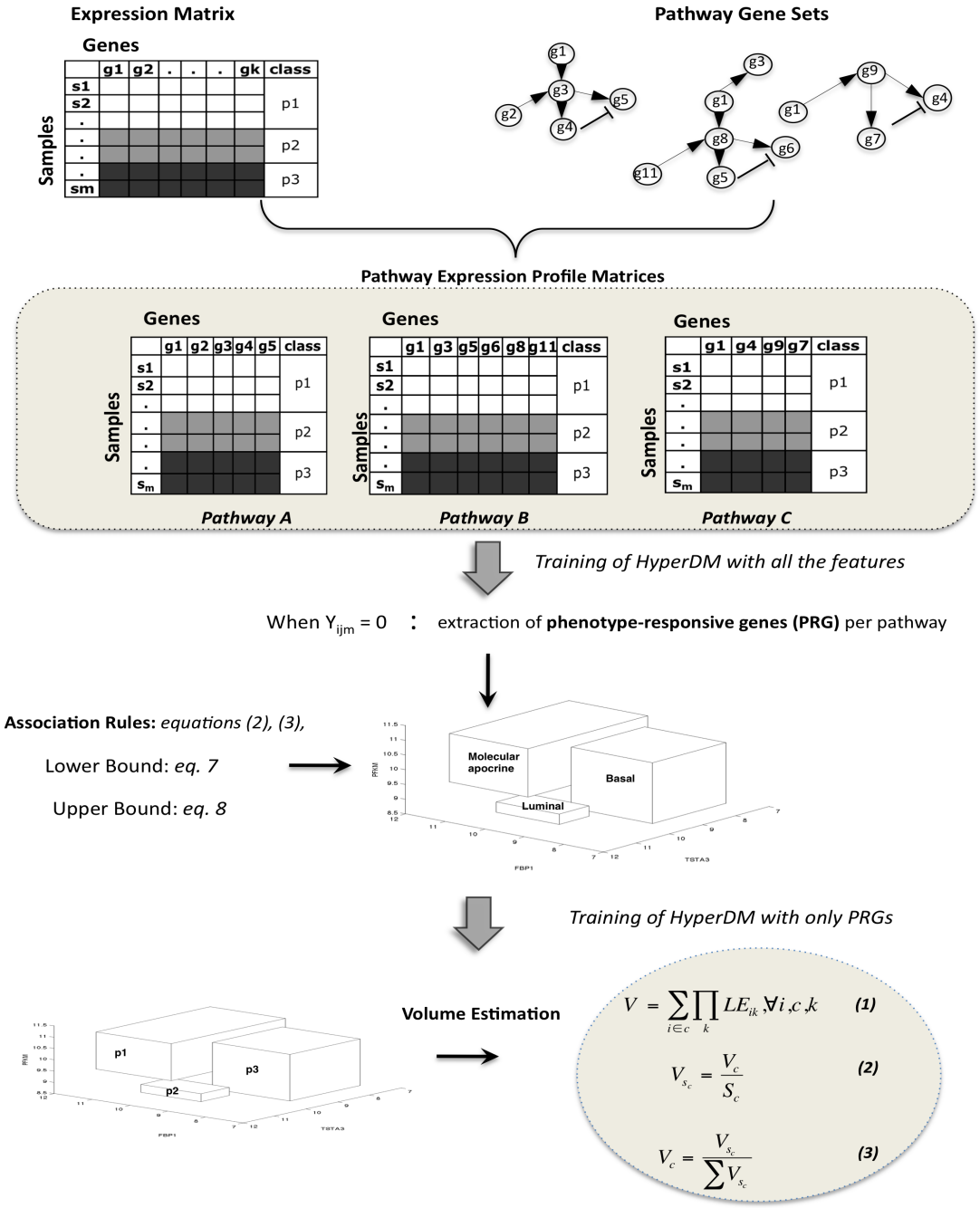

Figure 3.3: Schematic representation of the procedure to identify phenotype-responsive genes and extract association rules. For each pathway, corresponding expression values of participating genes were generated for all samples. HyperDM was applied to these pathway expression profile matrices and the association rules from hyper-boxes dimensions were considered to derive phenotype-responsive genes and infer the pathway activity.

locally in a parallel mode of 4 threads on a linux cluster consisting of 20 nodes (2 x Dual Core AMD Opteron(tm) Processor 280 running at 2.4 GHz and 8GB of DDR RAM memory, 2GB per core).

3.2.3 Identification of Phenotype-Responsive Genes (PRGs) Within a Pathway

A variant of HyperDM was also developed and applied by minimizing the attributes that contribute most on preventing the overlap of boxes that belong to different classes. For a given pathway, a new penalty term is added to the objective function to minimize the number of features that contribute to non-overlapping boxes. The new objective function minimizes both misclassified samples and the number of genes that guarantee the non-overlapping of two boxes i, j belonging to different classes.

Minimize:

$$\sum_{m}(1 - E_m) + \delta \sum_{k=1}^{K}(1 - Y_{ijk}) \tag{3.1}$$

where δ is a small positive number. Using the Y_{ijk} binary variable, a subset of genes is identified for each pathway, where each of these genes can differentiate two disease classes, for a particular gene expression spatial domain. We refer to this subset as "Phenotype-Responsive Genes" (PRGs) within a pathway representing the member genes that characterize the relevant phenotype. Furthermore, for the particular phenotype-responsive gene set (PRG) association rules are derived from the positioning and dimensions of the hyper-boxes, as described earlier (**Figure 3.3**). In this manner, we account for joint effects of multiple genes to the prediction and discrimination of disease phenotypes within each pathway. Note here, that the algorithm does not take into account the size of the pathways in order to produce the phenotype-responsive gene set (PRGs), but it uses the binary variable Y_{ijk}, as discussed above, to determine what is important for the specificity of each pathway.

3.2.4 Pathway specificity in disease phenotype

For a given pathway, the identified PRGs sets are further used to train another HyperDM model and the new dimensions were used to estimate the volume of each box in order to infer the specificity of each pathway to a specific phenotype. For each box, the length of any side is multiplied by itself and then the products of the boxes that correspond to the same class are summed up. In that way, the volume of each class per pathway is determined revealing how compact or concentrated the box is.

$$V = \sum_{i \in c} \prod_{k} LE_{ik}, \forall i, c, k \tag{3.2}$$

where LE_{ik} is the length of the box i on attribute k. In the case of highly unbalanced datasets in order to be unbiased on our results we normalize the volume by dividing each one with the number of samples that are involved in the relative class.

$$V_{s_c} = \frac{V_c}{S_c} \tag{3.3}$$

And finally we divide by the sum of the volumes:

$$V_c = \frac{V_{s_c}}{\sum V_{s_c}} \tag{3.4}$$

The higher the volume is the more disperse the data are, reflecting to less specificity of the pathway to the relative phenotype. An example is shown in **Figure 3.4** where equations 3.2, 3.3 and 3.4 applied to Farmer breast cancer dataset. Extension of this method is under development and comparison with other similar methods, such as weighted average of pathway genes, will be done.

3.2.5 Classification Evaluation

HyperDM was trained on the 186 pathway expression profiles for each of the six datasets using the genes involved in each pathway according to the MSigDB. To evaluate the predictive performance of HyperDM, other classification methods were also implemented to analyse the breast cancer and psoriasis datasets. Using

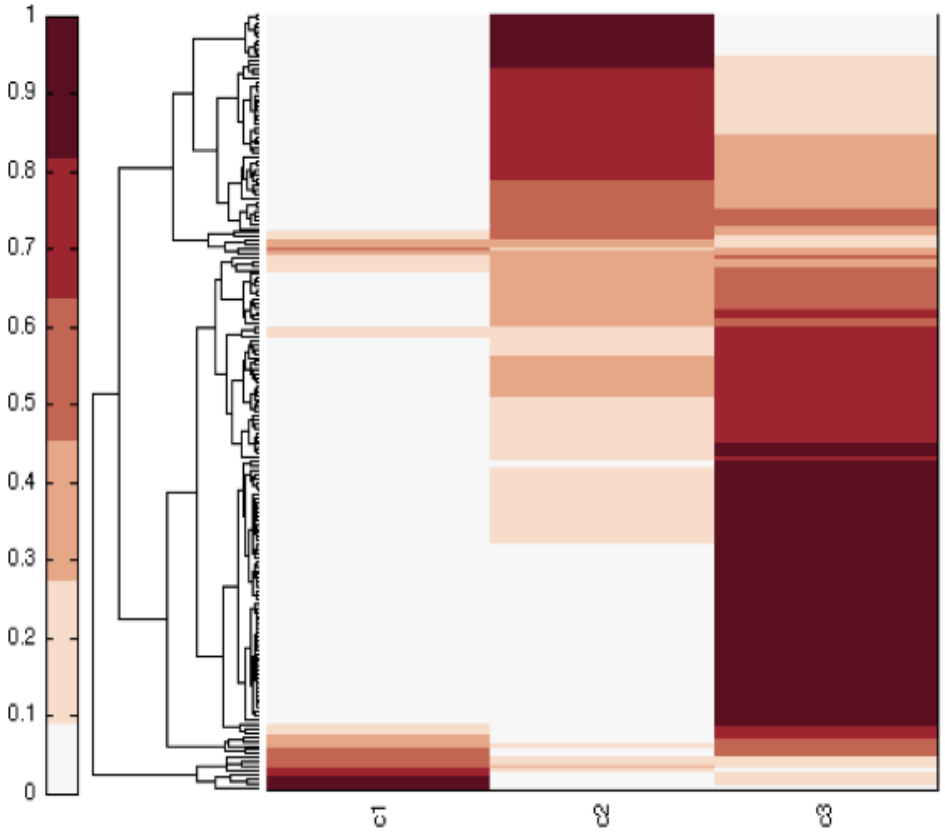

Figure 3.4: Pathway Specificity in disease phenotype using the volume metric derived for each pathway. X-axis represent the class nomenclature while Y-axis corresponds to each pathway.

the WEKA machine learning software (Frank *et al.* [2004]), Support Vector Machine (SVM) (Guyon *et al.* [2002]), Bagging and Random Forest (RF) (Breiman *et al.* [1984]) were applied as benchmarks and trained using the features involved in each of the pathways. Predictive power of classification methods was assessed through a training-testing scheme where 70% of the randomly extracted samples were used in the training set and the remainder 30% is used for testing. This scenario was repeated 50 times and the Mean Prediction Accuracy (MPA) with the Standard Deviation (StD) that reflects the average performance is reported. Algorithms were trained on the pathway expression matrices and parameters were set to the default. The performance and robustness of each classifier was evaluated based on the Signal-to-Noise Ratio (SNR) metric, which is given by:

$$SNR = 10 \times log(\frac{MPA^2}{StD^2}) \tag{3.5}$$

3.2.6 Hyper-box classification improves the discriminative power of pathway markers

Hyper-box classification was performed within gene sets corresponding to biochemical pathways as defined in KEGG. Here is presented the potential of the method to elucidate the joint effects of multiple genes within pathways to define particular disease states.

SVM outperforms for almost all the datasets, but HyperDM is promising for the small pathway sets (approx. up to 50 genes per pathway) while it still remains competitive in larger pathways. This is shown for two datasets (Farmer and Gudjonsson), in **Figrure 3.5**. It is worth noting that typical biochemical pathways usually comprise less than 50 genes, according to the KEGG database. The prediction accuracy of the model was compared to other classification methods and overall results are displayed in **Figure 3.6** for all the datasets. To estimate the robustness of the classification performance throughout the pathway in the different methods, the Signal-to-Noise Ratio (SNR) for each model was calculated. A robust method with low misclassification risk is expected to produce high SNR. According to (Table 3.6) HyperDM exhibits high SNR in some datasets indicating robust performance throughout.

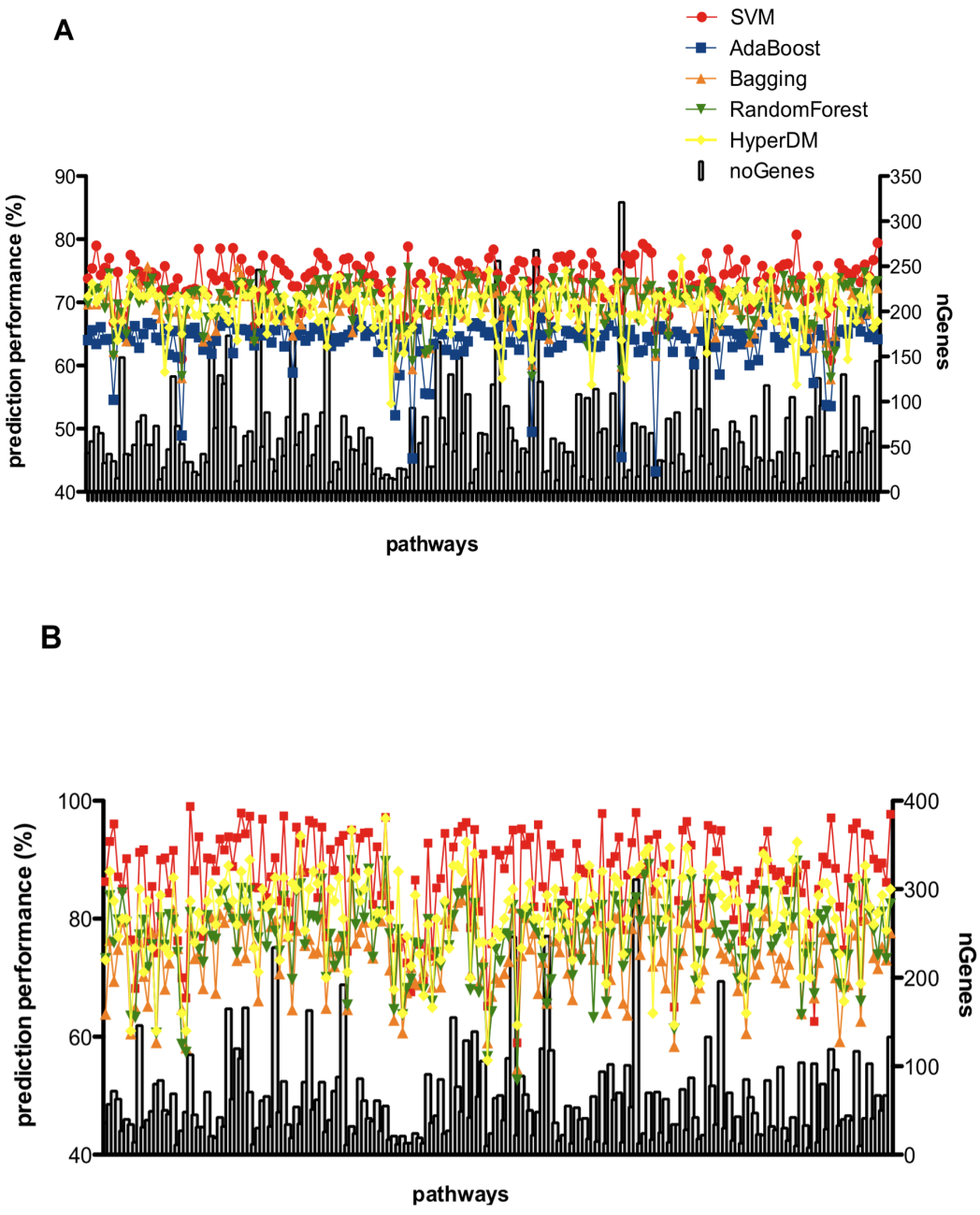

Figure 3.5: Prediction performance of HyperDM and comparison to other classification techniques applied in 186 pathways for (A) Farmer and (B) Gudjonsson datasets. Yellow line corresponds to HyperDM, red to SVM, orange to Bagging, blue to AdaBoost and green to Random Forest algorithm. Left y-axis represent the prediction performance, while right y-axis represent the number of genes included in each pathway. Pathways are shown in x-axis.

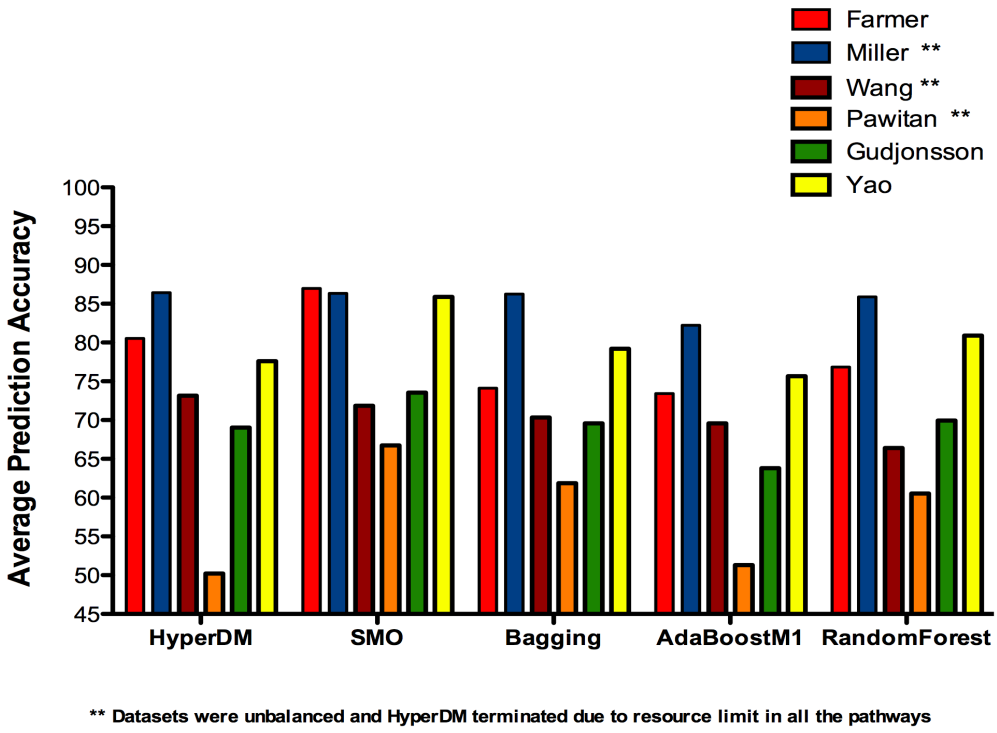

Figure 3.6: Comparison of different classification models across different disease datasets. Prediction performance was calculated as the average of the prediction performance estimated for each KEGG pathway separately

Table 3.6: Signal-to-Noise Ration (SNR) for the different algorithms applied to different datasets

Dataset	HyperDM.	SVM	AdaBoost	Bagging	RF
Farmer *et al.* [2005]	21.0974	21.0413	18.9177	19.2098	20.2008
Wang *et al.* [2005a]	17.054	18.321	13.321	14.432	18.032
Miller *et al.* [2005]	19.054	20.054	21.054	19.1234	20.032
Pawitan *et al.* [2005]	19.054	20.154	17.654	19.4356	20.234
Yao *et al.* [2008]	22.054	21.054	20.1254	19.023	20.456
Gudjonsson *et al.* [2010]	22.1342	22.123	20.0023	21.023)	20.043

From the biological perspective, this analysis yields results that are comparable to previous studies (Gatza *et al.* [2010]). For example, biochemical pathways related to energy metabolism infer high prediction accuracy among the basal, luminal and apocrine breast cancer classes. *Fructose and Mannose Metabolism, Glycolysis and Gluconeogenesis, Pentose Phosphate pathway, Pyruvate Metabolism, Neurodegenerative Disorders, JAK-STAT signaling pathway* and *MAPK signaling pathway* were identified as pathways with best classification accuracy, indicating a strong pathway-to-phenotype interaction. Note that energy metabolism is particularly important in cancerous cells, as it has been shown that several pathways for energy supply have been found to activate tumour cells with a metastatic potential (Schramm *et al.* [2010]). For example, it has been demonstrated that tumours have abnormal bioenergetics and subjects with cancer can show a systematic loss of energy involving the interaction of tumour glycolysis and gluconeogenesis (Perumal *et al.* [2005]). In addition, among pathways ranking on the top 30, many have been previously established as major contributors to the oncogenesis process, including *Apoptosis pathway, VEGF signaling pathway, Wnt signaling pathway, Toll-like receptor pathway* and *Cell Cycle*. Similar results were extracted also using the *GAGE* package from bioconductor which is a method for gene set or pathway analysis (Luo *et al.* [2009]).

3.2.7 Molecular signatures within pathways through phenotype-responsive genes (PRGs)

We extracted the maximum number of genes that separate the samples that belong to different classes. Within each pathway, genes that best differentiate two phenotype groups (PRGs) and have a joint impact on the disease classification, as described previously, are shown in Table 3.7.

For each pair of classes, the HyperDM model extracts at least one or more genes with the optimal discriminative power within the pathway. According to Table 3.7, for example, in Farmer dataset, within *Fructose and Mannose Metabolism* pathway HyperDM infers *PFKM, FBP1* and *TSTA3* genes out of 31 as the most differentially expressed among the three disease phenotypes. Interestingly, *FBP1, PFKM* and *TSTA3* were also among the top 10 selected genes identified by the WEKA using either SVM attribute selection, Correlation Feature Selection (CFS) or Random Forest (RF), providing further support to our results. *FBP1, PFKM* and *TSTA3* have also been shown recently to be up-regulated in cancer cells leading to ATP production (Schramm *et al.* [2010]). In addition, according to Breast Cancer Database (http://www.itb.cnr.it/breastcancer/) these genes have been previously identified to be susceptible to the disease. Gene-set signatures were also established for the other 10 top discriminative pathways. Table 3.7 shows various genes, previously known as promoters to the development of cancer, identified with HyperDM predictive model within different pathways. For instance, *MYC* gene has been previously shown to be prominent in basal-like breast cancer (Chandriani *et al.* [2009], Cancer Genome Atlas [2012]) and this is also confirmed from HyperDM analysis, where *MYC* best separates the basal group from the other two in JAK stat signalling pathway.

The proposed HyperDM can also provide rules on how samples from different classes are separated by the Phenotype-Responsive Genes (PRGs). It has been illustrated that the non-overlapping genes as output from the HyperDM method can well discriminate samples of different phenotypes. Each non-overlapping gene can separate samples from two different classes into two disjoint regions of expression values. Results regarding the inferred rules on how to separate different disease phenotypes can be seen at Table 3.8, for the first four top pathways in

Table 3.7: Gene signatures per pathway identified as differentially expressed among the different disease states in Farmer (breast cancer) and Yao (psoriasis) datasets

		Farmer Dataset	
		Phenotype-Responsive Genes (PGR)	
HyperDM top pathways	**Apocrine tumour/ basal tumour**	**Apocrine tumour/ luminal tumour**	**Basal tumour/ luminal tumour**
PENTOSE PHOSPHATE PATHWAY	ALDOA	PFKM	FBP1
PEROXISOME	AMACR,CROT, IDH1, IDH2,PEX11A,SCP2	IDH1	PEX11A
INSULIN SIGNALING PATHWAY	FASN	PRKAA1	FBP1
FRUCTOSE AND MANNOSE METABOLISM	TSTA3	PFKM	FBP1
GLYCOLYSIS GLUCONEOGENESIS	AKR1A1,ALDH3B2,LDHB	PGK1, PFKM	FBP1
PRIMARY BILE ACID BIOSYNTHESIS	ACOX2,AKR1D1, AMACR, SCP2,CYP39A1	HSD17B4	CYP7B1
SELENOAMINO ACID METABOLISM	PAPSS2, TRMT11	PAPSS2	TRMT11
PROPANOATE METABOLISM	LDHB	ACADM	ABAT
JAK STAT SIGNALING PATHWAY	PRLR,MYC, SOCS5, SPRED2	SPRY4, IL6ST	MYC
DRUG METABOLISM OTHER ENZYMES	CES1,UGT2B28	GMPS	GMPS
SULFUR METABOLISM	PAPSS2	PAPSS2	SUOX
		Yao Dataset	
		Phenotype-Responsive Genes (PGR)	
HyperDM top pathways	**Healthy control/ non-lesional skin**	**Healthy control/ lesional skin**	**non-lesional skin/ lesional skin**
B CELL RECEPTOR SIGNALING PATHWAY	JUN	JUN	BLNK,LYN
CYTOSOLIC DNA SENSING PATHWAY	NFKBIA	POLR3H	AIM2
NEUROTROPHIN SIGNALING PATHWAY	JUN	JUN	PIK3R5
TOLL LIKE RECEPTOR SIGNALING PATHWAY	JUN	MYD88	MYD88
CHEMOKINE SIGNALING PATHWAY	GNG10	PLCB1, PTK2	CXCR2
COLORECTAL CANCER	JUN	JUN	CCND1
RENAL CELL CARCINOMA	JUN	CREBBP	PIK3CA
ERBB SIGNALING PATHWAY	JUN	JUN, RPS6KB2	PIK3R5
T CELL RECEPTOR SIGNALING PATHWAY	JUN	JUN	IFNG
CYTOKINE CYTOKINE RECEPTOR INTERACTION	IL20RB	FLT1, IFNAR1,INHBB,CXCL9	CXCR2,IL4R

Farmer and Yao datasets, respectively. Taken as an example again the *Fructose and Mannose Metabolism pathway*, it is notable the different effect of same genes in the different classes. High expression of *TSTA3* (10.32<TSTA3 <11.24) and low expression of *PFKM* (8.72<PFKM <9.64) establish a signature specific for apocrine tumor, while low expression of *TSTA3* (8.1<TSTA3<10.13) and low expression of *FBP1* (7.27< FBP1<9.32) define the class as basal tumor. This play a major role in the final decision of therapeutic treatment.

Similarly, in psoriasis dataset, *JUN* gene is found to play a major role in the differentiation of healthy from psoriatic individuals (Table 3.7). Interestingly, *JUN* gene has previously shown to play a role in keratinocyte differentiation and be relevant in the pathogenesis of psoriasis (Rainer *et al.* [2005]). *CXCR2*, which is shown to better discriminate lesional cases from non-lesional one in *Cytokine-Cytokine Receptor Interaction* and *Chemokine signalling pathway*, has been previously detected to be over-expressed in psoriatic epidermis and to play a role in the pathogenesis of inflammatory responses (Saveria *et al.* [2007], Cataisson *et al.* [2006]).

Within each pathway, the HyperDM achieved to identify the genes that best differentiate two phenotypic groups (phenotype-responsive genes) and have a joint impact on the disease progression. This result suggests that in a pathway not all of the genes are transcriptional associated with the phenotype of interest.

3.2.8 Discussion

Classification of disease states in the context of pathway gene sets through an integer optimisation approach was demonstrated. A data from breast cancer and psoriasis expression profiles were used to illustrate the applicability of the method and a comprehensive comparison with other similar methodologies showed that HyperDM approach is competitive. A development of the original MILP formulation was done to achieve a minimum number of overlapping attributes that best differentiate the classes by preventing overlapping boxes. The findings evaluate a pathway-based approach through mathematical programming that lead to results closely tied with the biological mechanisms of complex diseases. However, a potential shortcoming of current pathway-based classifiers is that the pre-defined

Table 3.8: Association rules derived from the HyperDM model applied in Farmer (breast cancer) and Yao (psoriasis) datasets for the top pathways

		Farmer Dataset	
		Phenotype-Responsive Genes (PGR)	
HyperDM top pathways	**Apocrine tumour**	**Basal tumour**	**luminal tumour**
PENTOSE PHOSPHATE PATHWAY	ALDOA12.88<ALDOA<13.18	11.5<ALDOA<12.62	9.5<PFKM<11.14
	and 8.685< PFKM<9.015	and 7.27< FBP1<9.32	and 9.49<FBP1<11.48
PEROXISOME	[5.81<AMACR<7	[5.22<AMACR<5.68	10.38< IDH1<11.47
	or 9.63<CROT<10.93	or 8<CROT<8.89	and 7.53<PEX11A<9.06
	or 11.89<IDH1<12.26	or 9.95<IDH1<11.56	
	or 9.83<IDH2<10.87	or 8.38< IDH2<9.71	
	or 7.8<PEX11A<8.77	or 7.02<PEX11A<7.44	
	or 11.25<SCP2<11.82]	or 10.31<SCP2<11.1]	
	and 11.89<IDH1<12.26	and 7.02<PEX11A<7.44	
INSULIN SIGNALING PATHWAY	7.96<FASN<8.87	6.77<FASN<7.53	4.79< PRKAA1<5.22
	and 5.45<PRKAA1<6.59	and 7.27< FBP1<9.32	and 9.49<FBP1<11.48
FRUCTOSE AND MANNOSE METABOLISM	10.32<TSTA3<11.24	8.1<TSTA3<10.13	9.72<PFKM<11.14
	and 8.72<PFKM <9.64	and 7.27<FBP1<9.32	and 9.49<FBP1<11.48
		Yao Dataset	
		Phenotype-Responsive Genes (PGR)	
HyperDM top pathways	**Healthy control**	**non-lesional skin**	**lesional skin**
B CELL RECEPTOR SIGNALING PATHWAY	8.33<JUN<10.1	6.03<JUN<7.67	6<JUN<8.26
		and [5.83<BLNK<7.78	and [7.83<BLNK<9.19
		or 5.29<LYN<6.82]	or 6.99<LYN<8.67]
CYTOSOLIC DNA SENSING PATHWAY	11.62<NFKBIA<13.18	10.58<NFKBIA<11.61	5.9<POLR3H <6.36
	and 5.39<POLR3H<5.89	and 3.56<AIM2<4.37	and 4.44<AIM2<7.77
NEUROTROPHIN SIGNALING PATHWAY	8.33<JUN<10.1	6.03< JUN<7.67	6< JUN <8.26
		and 4.25<PIK3R5<4.67	and 4.69< PIK3R5<5.47
TOLL LIKE RECEPTOR SIGNALING PATHWAY	8.13< JUN <10.1	6.03< JUN<7.67	9.53< MYD88<10.47
	and 8.14< MYD88<9.45	and 8.42< MYD88<9.47	

set of genes making up a pathway may be derived from conditions irrelevant to the disease of interest. Future work can expand this approach on a more complete investigation of other pathways, while in-depth investigation of prediction rules can further aid understanding of the links between pathways or functional modules and disease phenotype.

3.2.9 Disclaimer

I am the sole contributor of the design, development, data collection and analysis of this work. The MILP model from Xu& Papageorgiou was previously developed, but further major development for application in disease classification was done by myself.

Chapter 4

Network Biology of Complex Diseases

Reductionist approaches are necessary in some cases, however current practise has indictaed that they may not be sufficient to capture important interactions between cellular components. Networks have been used to represent biological systems and to better understand disease mechanisms by discovering disease-associated genes. Network inference is another data-mining technique that can represent dependencies between variables in a dataset with a graph and can play an important role in medicine by understanding the pathophysiology of human diseases.

In this chapter, basic concepts in network biology will be discussed as well applications in complex diseases. Different types of network inference methods and brief description of molecular interaction networks will be given in **section 4.1**. In **section 4.2**, a novel application of network biology in the treatment of melanoma will be presented. Finally, in **section 4.3** a description and a small application of a novel gene network inference approach that has been developed using the Correlation Feature Selection algorithm (NetCFS) is introduced.

4.1 Basic concepts in network biology

A network ($G = (V, E)$) is modeled by a set of objects, vertices/nodes (V), that are connected to each other by links or edges (E). Nodes represent the elements of the system (e.g. genes, proteins, drugs, diseases) and the edges encode the interactions among those components (e.g. protein-protein interactions, disease-genes associations, gene-gene interactions, etc.). Depending on the nature of the interactions, networks can be directed or undirected. When vertices i and j are connected through an edge (i, j), then they are called neighbors. The total number of vertices (N) and edges E establishes the size of the network and identifies the number of distinct elements composing the system (Gautreau *et al.* [2009]). According to the connectivity patterns in the network, highly connected vertices represent a group, which may be functionally related and is named *network module*. In that way, the network is divided into communities and the modularity metric measures the granularity of this division. In biological systems, as in other systems, a network is an abstract representation of a graph that is defined by means of a an adjacency matrix ($n \times nmatrix$) representing which nodes are adjacent to which other nodes. **Figure 4.1** illustrates a typical biological network derived through gene expression data matrix.

The characterization of topological properties in a network is inferred through a variety of statistical concepts, such as *the degree distribution*, *the network diameter*, *the clustering coefficient*, *the shortest path length* and *the betweenness*.

The *degree* of a node (or connectivity) refers to the number of links that node has to the other nodes. The probability that a node has exactly k links, $P(k)$, is obtained by estimating the number of nodes $N(k)$ with $k = 1, 2, n$ links and dividing by the total number of nodes N. This is called *degree distribution* and reflects the different types of networks. When the probability distribution of the number of edges per node follows a Poisson distribution and there are no highly connected nodes (most nodes have roughly the same numbers of links, thus there are no hubs), the network is called random. Otherwise, when there is a number of nodes in the network that are highly connected (hubs) and the majority of the nodes have small number of connections, then the network follows a power-law distribution and is called scale-free (Vazquez *et al.* [2004], Barabasi [2005]).

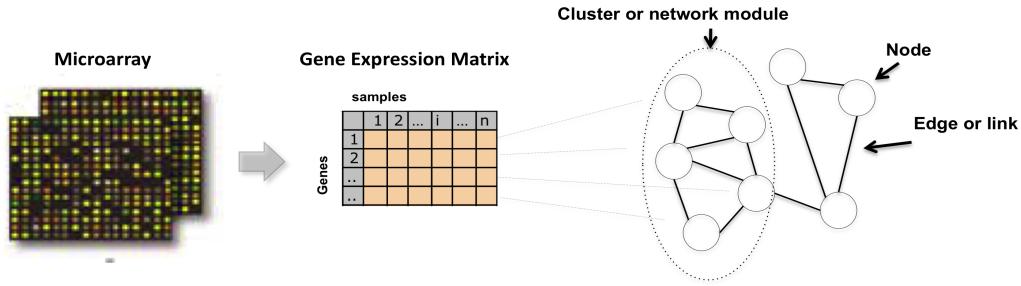

Figure 4.1: A typical biological network of gene-gene interactions derived from a microarray dataset. Visualisation of the basic concepts of a network/graph, as described in the **section 4.1**.

Most biological networks are scale-free and hub components play a major role in uncovering the underlying molecular mechanisms of disease, as will also be seen in the applications described in **section 4.2**.

The *clustering coefficient* of a node is a measure of the connection density around a node and is given as the proportion of the observed connections between the neighbors of the node. *Clustering coefficient* is given by the formula 4.1:

$$C_i = \frac{2n_i}{k_i(k_i - 1)} \tag{4.1}$$

where n_i is the number of edges connecting the k_i neighbors of node i and $\frac{k_i(k_i-1)}{2}$ is the total number of possible links.

The shortest series of connected edges that separates two nodes is defined as the *shortest path* between these nodes. The *network diameter* is the maximum length of all shortest paths between all pairs of nodes in a network. If a network is disconnected, its diameter is the maximum of all diameters of its connected components. The *betweenness* of a node quantifies the number of non-redundant shortest paths passing through the node. Note that the *network diameter* and *the shortest path length* distribution may indicate small-world properties of the analyzed network (Barabasi & Oltvai [2004],).

4.1.1 Types of biological networks

There are many different types of biological networks representing various biological systems; protein-protein interactions, metabolic networks and gene regulatory networks are mainly used in this thesis and discussed further here.

Protein-protein Interaction Networks

Proteins are important macromolecules of life and facilitate most biological processes in a cell. When two or more proteins bind or act together to undertake praticular biological functions at cellular and systems level, they interact and generate proteome-wide physical connections (De Las Rivas & Fontanillo [2010]). Those physical connections are essential to achieve a comprehensive knowledge of protein interactomes that will lead to a further understanding of diseases, along with the development of therapeutic targets.

Protein-protein interactions (PPIs) are available from multiple sources and

there are various databases that report known human protein-protein interactions (PPIs) (Table 4.1). Note here, that due to literature curation of protein interaction data, the public databases may not agree in the interactions or proteins included due to the divergent curation policies across databases (Turinsky *et al.* [2010]). PPIs are represented as networks, where nodes are proteins and the edges correspond to proteins that either physically bind to each other or share similar functions, for example participation in the same biochemical pathway (Yook *et al.* [2004]).

Table 4.1: Human protein-protein interaction databases

Databases	**Number of proteins**	**Number of Interactions**	**URL**
HPRD	30047	39233	http://hprd.org/
CCSB interactome	N/A	13944	http://interactome.dfci.harvard.edu/
DIP	3337	4540	http://dip.doe-mbi.ucla.edu
BioGRID	14519	62821	http://www.thebiogrid.org
MINT	8625	32249	http://mint.bio.uniroma2.it/mint

Metabolic Networks

A series of biochemical reactions occurring within the cell are organized into metabolic pathways and are connected by their intermediates. In each pathway, enzymes catalyze these biochemical reactions, transform one metabolite into another and further structure a metabolic network. In such a network that models the metabolism, nodes correspond to metabolites and enzymes, and edges are biochemical reactions that convert one metabolite into another. In that way, cellular metabolism is modeled as a large-scale organizational picture in a network-based representation (Schuster *et al.* [2000], Stelling *et al.* [2002]).

Databases with networks of biochemical reactions are available for many organisms. Such are: Kyoto Encyclopedia of Genes and Genomes (KEGG)(Kanehisa *et al.* [2004]), EcoCyc (Keseler *et al.* [2011]), BioCyc (Caspi *et al.* [2012]).

Gene Regulatory Networks (GRNs)

With the availability of complete genome sequences, the regulation of expression of the sets of genes is encoded by Gene Regulatory Networks (GRNs). Genes are

connected with physical or regulatory interactions and represent the causality of developmental processes. Genes can be viewed as nodes in the network, while interactions between genes depict the edges (Davidson & Levin [2005], Levine & Davidson [2005]).

The characterization and understanding of gene networks from high-throughput microarray expression data reveal global principles of gene regulation. A challenging step is how to infer those networks in order to deduce the functions for each gene and to better understand the behavior of the complex biological systems, from gene to signaling pathway, cell or tissue level. Thereby, gene network inference is a fundamental task and several methods have been proposed that are described in more detail in **section 4.1.2**.

4.1.2 Inference methods in biological networks

The reconstruction of biological networks from high-throughput data received enormous interest the last years. Mathematics, multivariate statistics and information science methods have been developed and applied to reconstruct gene regulatory networks and used to better understanding these systems (Bansal *et al.* [2007], Beer & Tavazoie [2004], Gardner & Faith [2005], Prakash & Tompa [2005], Ambesi-Impiombato *et al.* [2006]).

Supervised and unsupervised learning are the most widely used methods in gene network reconstruction, as described in **Chapter 2**. Unsupervised is based on correlation and on mutual information, while supervised require knowledge of regulatory interactions to be used as a training set (Madhamshettiwar *et al.* [2012]). Pearson Correlation Coefficient (PCC) is used to define co-expression of genes across the various experiments. In a co-expression network, the nodes represent the genes and the edges represent the degree of similarity in the expression profiles of these genes, according to the equation 4.2.

$$PCC_{ij} = \frac{\sum_{k=1}^{M}(x_i(k)x_j(k))}{\sqrt{(\sum_{k=1}^{M} x_i^2(k) \sum_{k=1}^{M} x_j^2(k))}} \tag{4.2}$$

Mutual Information (MI) is used to measure the degree of indepedence between two genes (equation 4.4).

$$MI_{ij} = H_i + H_j - H_{ij} \tag{4.3}$$

where H is the entropy and defined:

$$H_i = -\sum_{k=1}^{n} p(x_k) log(p(x_k)) \tag{4.4}$$

where $p(x_k)$ is the probability distribution of random variable x that takes on values k.

In practical application, has been shown that MI and Pearson correlation may yield almost identical results (Steuer *et al.* [2002]).

Several other approaches such as Bayesian networks (Friedman *et al.* [2000], Yu *et al.* [2004]), and ordinary differential equations (ODEs) (di Bernardo *et al.* [2005], Bansal *et al.* [2006]) were also used successfully to infer gene networks. Gene network inference algorithms have been used in a wide range for microarray data analysis and recently have started to become more "integrative" by exploiting, proteinprotein interaction data, sequence data, genomic data and others (Workman *et al.* [2006]).

4.2 Protein-protein interaction networks exhibit novel molecular signatures in melanoma treatment

Network-based approaches are widely used in the context of health, disease and biomarker discovery. For instance, different types of cancer phenotypes were investigated through the topological properties of networks (Chuang *et al.* [2007], Xu *et al.* [2008]); pathways were identified and local structures were characterized for cardiovascular and infectious diseases (Sharan *et al.* [2005], Camargo & Azuaje [2007]). The promise of network biology is to integrate diverse biomedical characterizations into detailed models of underlying mechanisms and therapies through suitable computational and mathematical formalisms.

Malignant melanoma is the most lethal form of skin cancer with an inci-

dence of 3.1 per 100,000 and an estimate of 150,000 - 200,000 cases diagnosed every year worldwide (Ferlay *et al.* [2010]). Melanoma is considered the archetypal immunogenic tumour, because of its associations of melanoma lesions with considerable immune cell infiltrates (Ramirez-Montagut *et al.* [2003], Tjin *et al.* [2011]). Temporal anti-tumoral antibody and memory B cell responses in lesions and in patient circulation have also been reported for a number of years (Gilbert, Karagiannis et al. 2011) (Selim, Burchette et al. 2011). However, prognostic evaluations in melanoma remain challenging and novel diagnostic and prognostic biomarkers and tools are needed to improve prediction of clinical outcomes, guide important clinical decisions, inform choice of suitable treatment regimes and identify patients who may benefit from particular therapies (Balch & Soong [2008]).

Most antibodies for cancer therapy in clinical use belong to the immunoglobulin G class (IgG), because IgG was reported to be the best complement activator. However, immunoglobulin E (IgE) antibodies function through high-affinity Fc receptors and recent studies suggest that IgE antibodies are more effective than the corresponding IgG antibodies in eliciting anti-tumor immune responses. In this study, was examined if and how IgE antibody may harbor functional properties against cancer cells. First, IgG1 and IgE antibodies were engineered of the same specificity against the human high molecular weight melanoma associated antigen (HMW-MAA). In vitro investigations revealed that IgG1 and IgE antibodies induced significant tumor cell death. Then, comparison of the therapeutic efficacy of these antibodies in tumor-bearing NOD/SCID/IL-2 receptor chain-/- mice engrafted with human lymphocytes (Figure 4.2) was done. Therefore, by activating different families of Fc receptors on immune effector cells, we demonstrate that: (i) different antibody classes may play a functional role against cancer cells and (ii) IgE antibodies are superior compared to IgG1 ($p<0.05$) and non-specific antibody controls ($p<0.001$).

The translational value of the resulting knowledge lies in synergistically combining bioinformatics and lab-based research to unravel biomarkers and pathways that can be targeted using novel therapies but also in evaluating key immunological pathways and components in melanoma as prognostic indicators or predictors of clinical responses to treatments. Herein, a systems analysis to identify the

effect of IgE and IgG antibodies in the tumor was performed. The objective is to investigate the genes and pathways that affect tumors when treated with either antibody (IgG or IgE) to assess the functional differences and similarities.

4.2.1 Data Sources

Seven NSG mice per group were used in a model of antitumor alloimmunity **Figure 4.2**. NSG mice were injected subcutaneously with 5×10^5 A375 tumor cells on day 0. On day 5 mice were injected intravenously with 10×10^6 human peripheral blood lymphocytes (PBLs), which were isolated from a cohort of 8 healthy volunteers **Figure 4.2A**. Subsequent injections of antibody treatments were given 3 times on days 12, 18, and 25 at doses of 10 mg/kg each in 150 l baseline (PBS). A control group was treated with 10×10^6 human PBLs on day 5 and injected with 150 l of PBS on days 12, 18, and 25. Tumor growth was monitored and measured using calipers. Tumor size (mm^3) was calculated using the following formula: $mm^3 = d^2 \times (D/2)$, where d stands for the small diameter of tumor and D stands for the large diameter of tumor. Experiments and data acquisition were adapted by Dr. Panos Karagiannis, St. John's Institute King's College London.

After assessing for RNA integrity, Human Ref-6 and Ref-12 BeadChips (Illumina, Ambion) were used to generate RNA expression data from tumors as per manufacturers instructions. Beadchips were scanned with a BeadArray Reader (Illumina, San Diego, CA). Each array contains over 50000 probe sets representing approximately 40000 human genes. Data were background-subtracted then log-transformed and quantile-normalized. The expression levels were compared by limma method an empirical Bayesian method with a moderated t- statistic implemented in Bioconductor and the Benjamini-Hochberg method was used to correct for multiple testing. A gene was considered as differentially expressed if the q-value was below 0.05 and fold change (FC) more than 1.2 and less than -1.2.

Treatment groups IgE (MAAIgE) and IgG (MAAIgG) were compared versus baseline (saline control, PBS). The analysis started with seven samples per group, but due to low quality of some experimental results, some sample had to be

removed. Comparison of the groups was followed using 4 samples for PBS group, 6 samples for IgE group and 2 samples for IgG group. For the IgE treatment group, 1045 genes were identified as significantly modulated; 1006 genes were over expressed in IgE (MAAIgE) and 39 under-expressed. For the IgG (MAAIgG) group were found 559 genes to be significantly modulated; 547 over expressed in IgG (MAAIgG) and 12 under-expressed.

Interestingly, as shown in **Figure 4.3A**, 510 genes found to have common effects and were over-expressed in both IgE and IgG samples compared to the baseline. 496 genes were prominent to IgE and only 37 were specific to IgG, showing a notable impact of IgE compared to IgG in eliciting anti-tumoral immune responses, including genes related to immune response as descrived in **section 4.2.2**. **Figure 4.3B** shows the low number of under-expressed genes found to be common in IgE and IgG (only 8 genes common; 31 genes specific to IgE and 4 genes specific to IgG)

4.2.2 Characterization of differentially regulated genes in samples injected with IgE and IgG antibodies

To better understand the functional role of the genes that were dysregulated in IgE and IgG compared to baseline samples, a comprehensive functional enrichment analysis was further implemented using the Database for Annotation, Visualisation and Integrated Discovery (DAVID) v6.7 [1]. Gene Ontology (GO) terms and over-represented pathways of the genes involved in the two groups were assessed. Over-expressed genes were enriched in energy related functions as well in energy related and signalling pathways (e.g. oxidative phosphorylation, glycolisis/gluconeogenesis pathway, integration of energy metabolism, Signaling by Wnt etc.). In contrast, down-expressed common genes included enrichment in binding related functions and trascription regulator activity. Genes that were specific in IgE, were found to be enriched mainly in metabolic pathways as well JAK-STAT and chemokine signalling pathways, whereas genes that were specific to IgG antibody were enriched in Insulin signalling pathway. An over-representation of immune related processes was also discovered indicating the significance of IgE

[1] http://david.abcc.ncifcrf.gov/

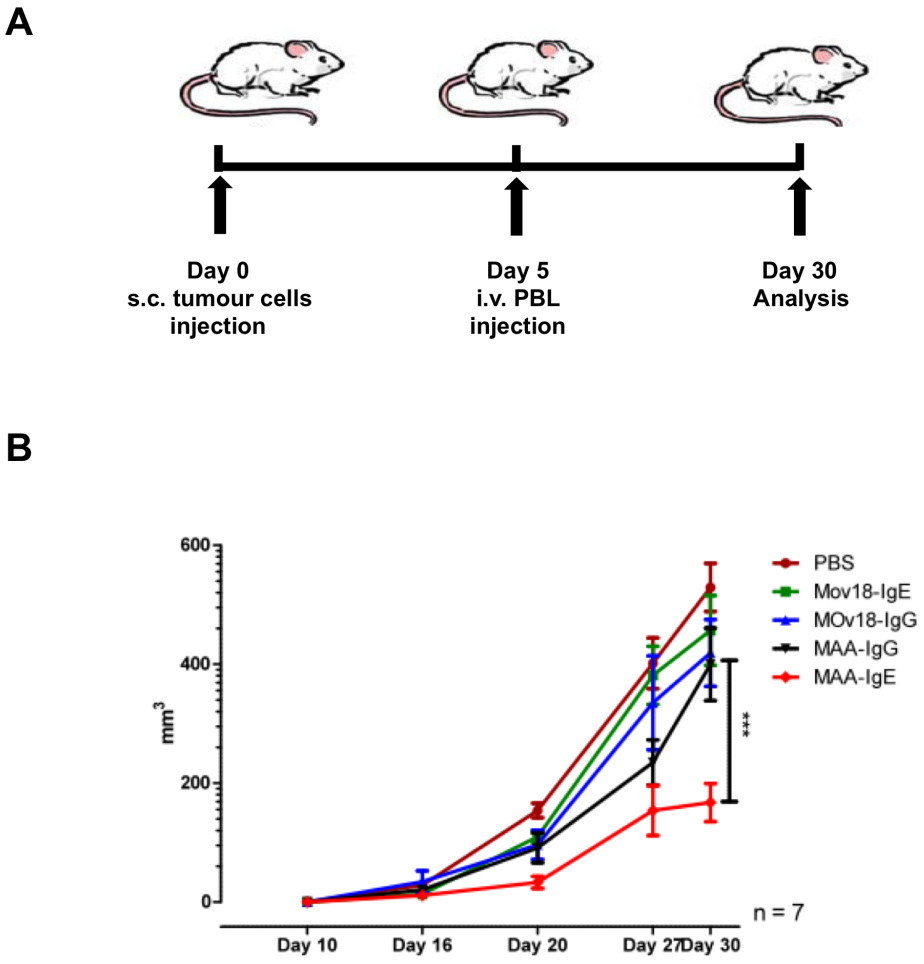

Figure 4.2: Humanized mice models used to identify the effects of IgE and IgG antibodies in tumor. (A) Schematic flowchart of the experiments in Humanized mice models of melanoma antitumor alloimmunity (B) effects of IgE and IgG in tumor growth comparing to non-specific antibody and baseline (PBS) (n = 7 mice per group; mean SEM tumor volume in mm^3
).*P $<$ 0.05, ***P $<$0.001, 2-way ANOVA with Bonferroni post-hoc test. Data are representative of 2 experiments.

in inducing anti-tumoral immune responses.

4.2.3 Protein-Protein interaction networks in IgE and IgG

Five broadly used databases of known human protein-protein interactions were analysed for their screening and inclusion criteria of interactions as well as the size of the databases (as shown in Table 4.1). BioGrid and HPRD are the largest databases of human interactions and therefore, a protein-protein interaction network from the consensus interactions of those databases was generated.

The initial interaction network contained 12719 interactions documented in both BioGrid and HPRD databases; however our analysis focused on the differentially expressed genes identified in IgE and IgG in our experiments. There were 1172 interactions that link 1110 nodes in the commonly expressed genes as explained above (**Figure 4.4A**). 1226 protein-protein interactions with 1162 nodes were identified in the specific to IgE genes (**Figure 4.4B**), whereas there were only 65 interactions linking 80 nodes in the protein-protein interaction network derived from IgG specific genes (**Figure 4.4C**). To further examine the association of the differentially expressed genes with the network, the up- and downregulated genes were mapped onto the network and colored them by red and green, respectively (**Figure 4.4**). In these networks, hub proteins seem to play significant role since in the majority are observed as up-regulated in the experiments and are communicating with many other up-regulated proteins being over-represented.

To evaluate the statistical significance of the IgE, IgG specific and common networks, randomized networks were generated by reshuffling edges in the original network while preserving the degrees of the respective nodes. The number of rewiring edges taken for each model was $4 \times (numberofedges)$ and the procedure was repeated 300 times. The simulation showed that on average, the three random networks contained small number of nodes proving the statistical significance of the identified hubs in each of the real networks. In other words, nodes in the randomised networks exhibit low number of connections, low degree nodes.

Therefore, as shown in **Figure 4.5**, the analysis exhibited that the network follow a power law distribution, which is consistent with a "scale free network"

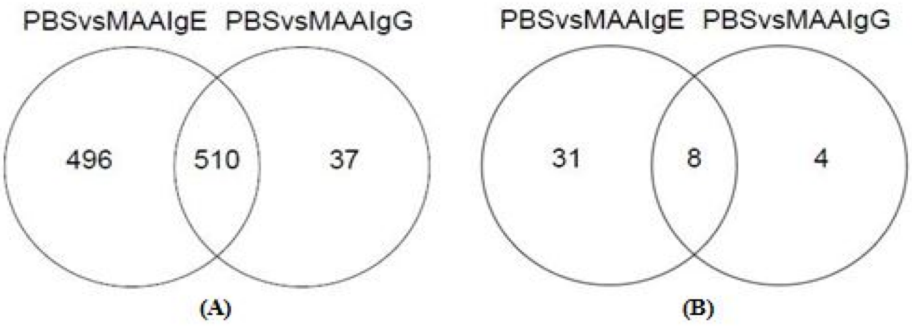

Figure 4.3: Venn diagram representing the number common and different genes being (A) over-expressed and (B) under-expressed in IgE and IgG samples comparing to PBS (baseline).

reported for most biological networks analyzed to date (Barabasi & Oltvai [2004]; see **section 4.1**). It is also notable that the R-squared values, which range between 0.7-0.9 also supported the view that the networks are scale-free (**Figure 4.5**). The R-squared value for the degree distribution is highest for the network specific to IgE (**Figure 4.5B**) indicating higher correlation and a stronger linear relationship between the data variables.

Furthermore, in order to examine the network topology, the topological parameters, like average clustering coefficient, average degree, network diameter and density, were assessed (Table 4.2). The protein-protein interaction network that derived from IgE specific genes has the largest diameter and the clustering coefficient is also higher (19 and 0.018, respectively) compared to the others. In addition, the density metric for this network appears to be small displaying the compactness of the graph (Table 4.2). This highlight a structured graph that also depicts the possibility that local network topology might be driven by biological relationships between proteins and illustrates the possible biological effect of IgE in the pathophysiology of melanoma against tumor growth. Intriguingly, this hierarchical architecture of our network suggests also the existence of small, highly connected, topological modules with discrete functions as shown in the next section (Ravasz *et al.* [2002]).

Table 4.2: Topological parameters of protein-protein interaction networks

Networks	Nodes	Edges	Average Degree	Diameter	Density	Clustering Coefficient
Common	1110	1172	2.111	17	0.0019	0.015
IgE specific	1162	1226	2.110	19	0.0018	0.018
IgE specific	80	65	1.625	4	0.0206	0

4.2.4 Discovery of biologically functional modules in PPI networks

Louvain method was implemented for identifying communities in our networks. The method is a greedy optimization approach that first looks for small communities by optimizing modularity locally and then iteratively aggregates nodes

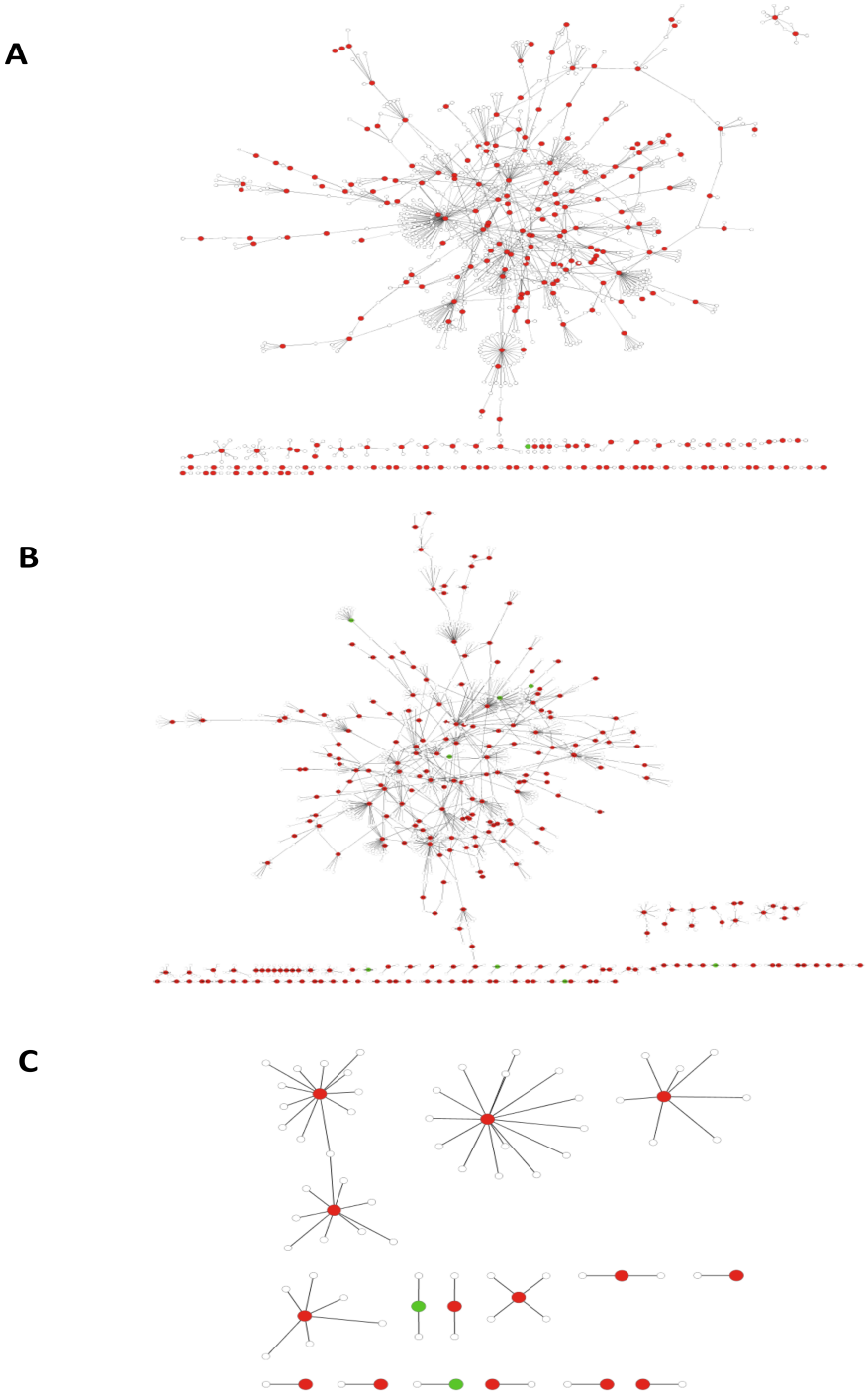

Figure 4.4: Protein-protein interaction networks for (A) the common up and down regulated genes, (B) up and down-regulated genes specific to IgE comparing to baseline samples and (C) up and down-regulated genes specific to IgG comparing to baseline samples. Red spots correspond to over-expressed genes while green spots correspond to under-expressed ones.

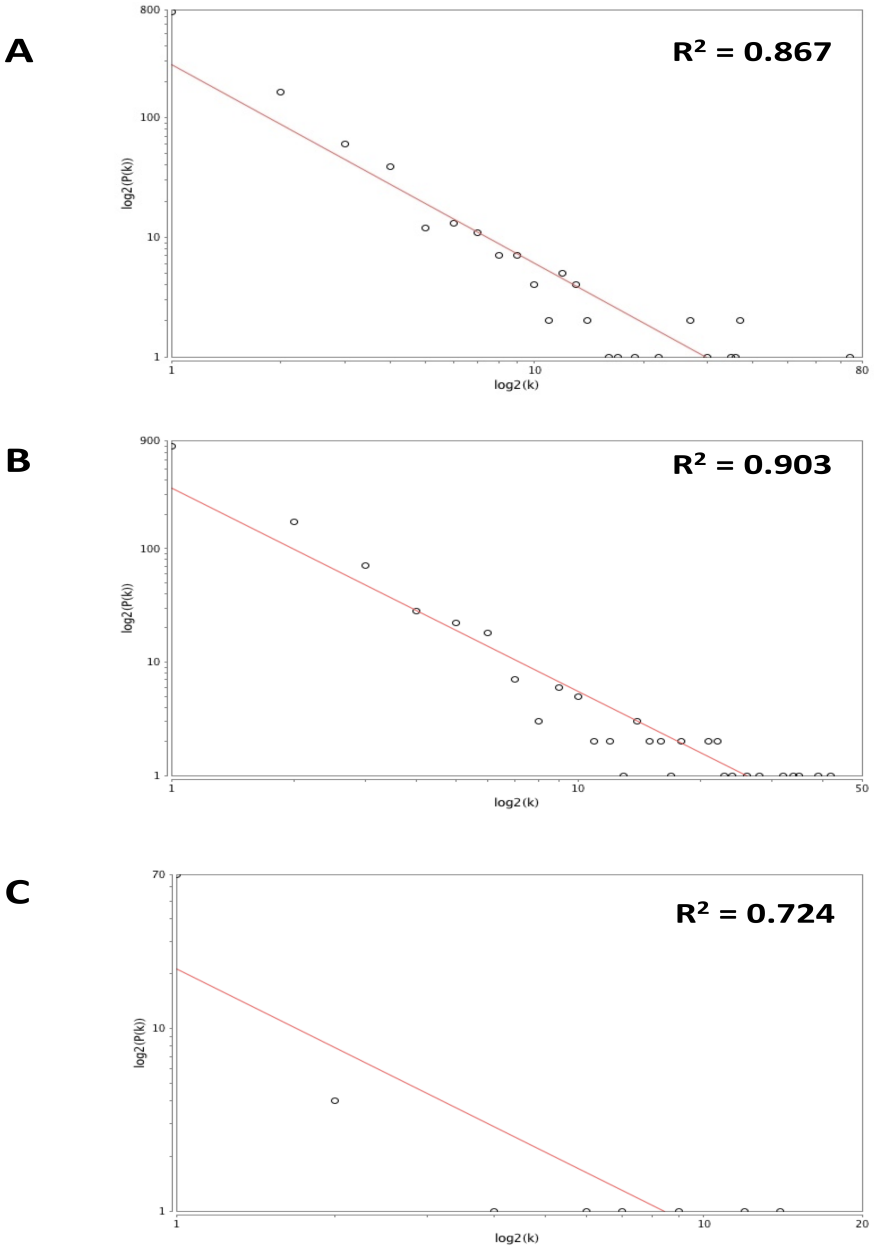

Figure 4.5: Power-law node degree distribution for the three protein-protein interaction networks. (A) common up and down regulated genes, (B) up and down-regulated genes specific to IgE comparing to baseline samples and (C) up and down-regulated genes specific to IgG comparing to baseline samples. A scale free topology is shown where there is a decrease in degree distribution with an increase in the number of links.

belonging to the same community and builds a new network whose nodes are the communities (Subelj & Bajec [2011]). In the IgE specific PPI network, 128 modules were found (**Figure 4.6A**) and further the functional enrichment of the largest clusters was examined.

Two of the largest groups that were constructed are shown in **Figure 4.6B**, which is consisting of 85 nodes of which the 16 found to be over or under expressed in the experiments. These clusters are enriched in Fc receptor signalling (i.e. FcRI, FcR) therefore, validating the hypothesis that anti-tumour activity is immune mediated; furhter investigation of their downstream functions was also done. Interestingly, the kinase PTK2 (p-value=0.01) which is connected to Fc signalling is up-regulated in IgE treated mice and according to **Figure 4.6B** is connected with GSK3B which is directly interacting with NFKB1. Therefore, this kinase is indirectly connected with NFKB1 and does regulate the IgE mediated tumour killing in vitro. Note here, that PTK2 is also indirectly connected with SOS1 (which is under-expressed in our experiments and involved in Fc epsilon RI signalling pathway) via NCK2, CRK and FYN. In vitro functional assays validated that inhibition of PTK2 in monocytes reduces their potential to induce antibody depended cytotoxicity (**Figure 4.7**). Analysing the phosphorylation of PTK2 from primary monocytes that were used in an ADCC/ADCP assay to induce tumour death, was found that IgE mediated cell death was inducing the phosphorylation of PTK2. Further blocking PTK2 activity in primary monocytes with the kinase inhibitor PF-431396 prior to the ADCC/ADCP assay inhibited IgE-mediated tumour cell killing significantly.

4.2.5 Discussion

Network biology is important to gain an insight into biological responses (Sharan *et al.* [2005]). Here, known protein-protein interactions from public databases were integrated with gene expression profiles gathered from a humanised mice models in melanoma. Combining information from changes in gene expression with topological criteria from the analysis of interaction networks, the effects and the functional roles of immunoglobulin antibody classes were investigated against cancer cells.

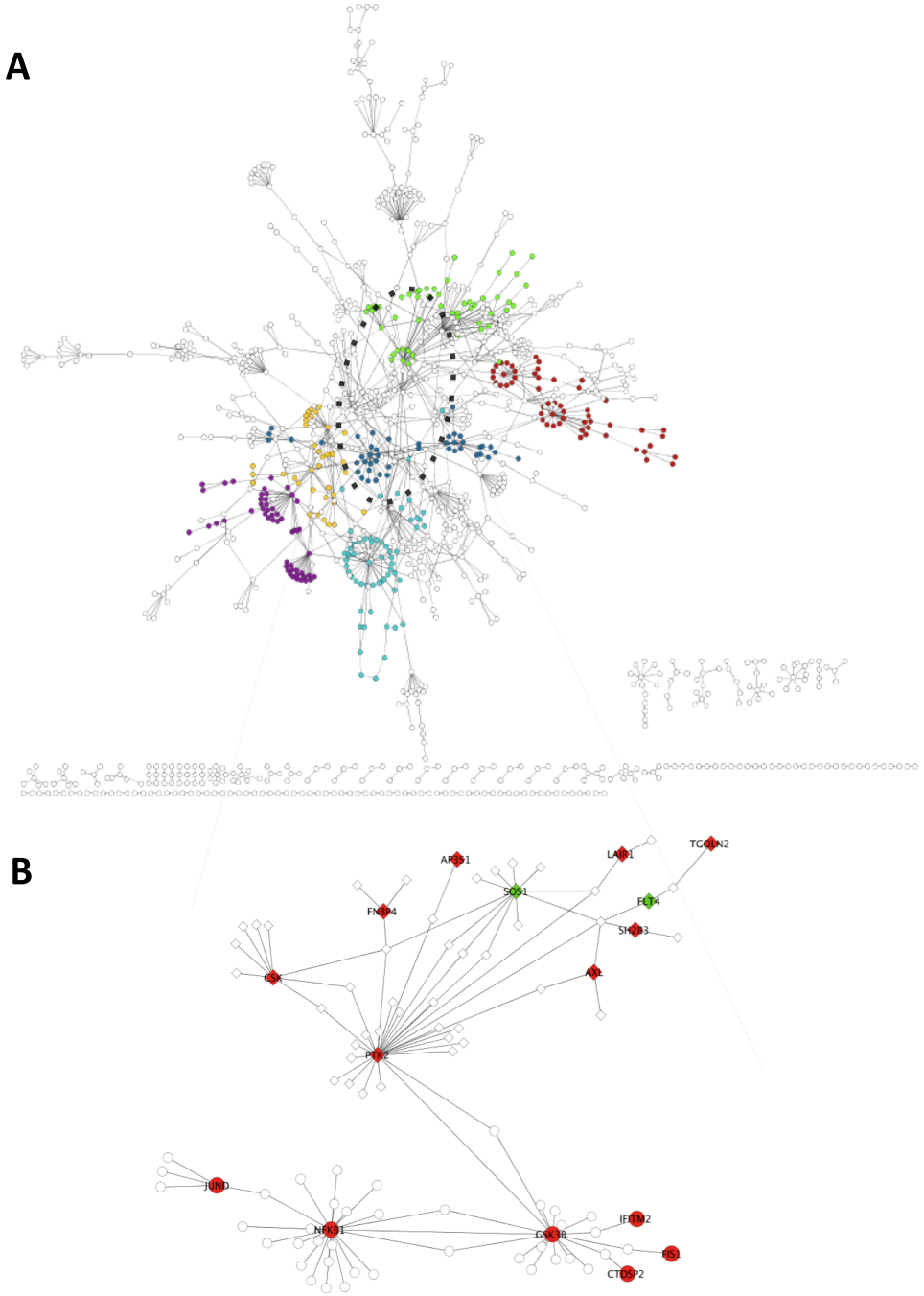

Figure 4.6: Clustering of Protein-Protein Interaction IgE network. (A) IgE specific protein-protein interaction network separated into modules (colored spots). Different color describe different module. Six of the largest modules are visualised. (B) Modules related to Fc receptor pathway and other immune-related functions. Red spots correspond to over-expressed genes while green spots correspond to under-expressed ones. Diamond shaped spots are involved in the first module while circle shaped spots consist the second module.

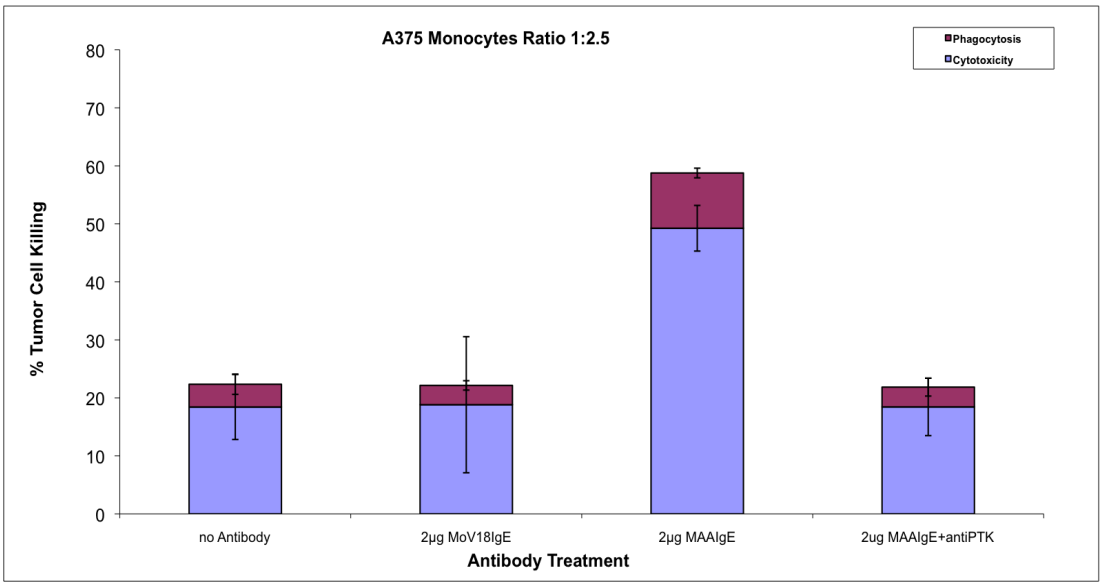

Figure 4.7: Inhibition of PTK2 in monocytes reduces their potential to induce antibody depended cytotoxicity. Tumor cells killing mediated by IgE was inhibited by the specific PTK2 kinase inhibitor PF-431396, indicating the relevance of this kinase in the downstream signaling of IgE.

The superiority of IgE compared to IgG was computationally validated according to the graph topological properties and subtle differences in the architecture of the networks. A topological and functional analysis of IgE network was followed and the observations indicate interesting functional modules, biomarkers and novel relationships of proteins whose gene expression is impaired in the experiments. In order to further confirm the relevance of putative targets and molecular interactions found by the network analysis further experimental validation is performing at Guy's Hospital - St. Jonh's Institute, with in vivo and in vitro models to study their underlying biological function.

4.3 Gene Network inference through Correlation based Feature Selection (NetCFS)

The inference of gene interactions from biochemical experiments is a crucial problem in the post-genomic era that received much attention especially in the field of

network medicine and translational bioinformatics (Marbach *et al.* [2010]). Several computational techniques, such as machine learning approaches, have been widely used to generate gene expression networks (Lee & Tzou [2009], Marbach *et al.* [2010]). Models based on statistical analysis, such as correlation coefficient and mutual information, were used successfully for the inference of gene regulatory networks (Butte *et al.* [2000], Eisen *et al.* [1998]). Similarly, supervised classification approaches have been used to reconstruct gene networks in functional genomics. Decision trees have been employed to predict changes in the expression of a gene based on the changes in the expression of the other genes (Soinov *et al.* [2003]). More recently, Random Forest (RF) was used to construct regulatory gene networks (Huynh-Thu *et al.* [2010]). The basic idea is to pick a gene as a predictor, and build a random forest classifier for that gene, using the remaining features. Feature selection from the Random Forest returns a ranking for all the genes according to how informative they are to the predictor, as discussed in **Chapter 2**. This process is repeated for every gene to determine subset of genes whose expression is predictive of the target gene. The main challenge with this approach is that feature selection returns a ranked list of genes, requiring a threshold for choosing the best subset of features. Therefore, the problem of choosing an optimal confidence threshold induces bias in this approach. An implementation of this algorithm, named *GENIE3*, using the programming languages R and Matlab has been reported.

The objective is to find a way of solving this threshold problem, using a similar concept as GENIE3 to generate gene expression networks. The method can contribute to this endeavor by using Correlation based Feature Selection (CFS) technique to induce the target features used to decompose the problem into g different regression problems. In that way, the solution will discard the threshold constraint and exploit a subset of features chosen by the algorithm. The principle idea of our method is based on the identification of g correlated features to generate g sub-problems. In the next section, we will introduce a novel inference algorithm, *NetCFS*, with its components. Also, an application will be described to present the mechanisms of the algorithm.

4.3.1 NetCFS algorithm

In **Chapter 3** was demonstrated the importance of CFS algorithm to extract subset of genes that are uncorrelated with each other but highly correlated with the class. CFS works perfectly with numeric classes and is simple by producing the results relatively quickly, compared to other feature selection techniques. According to the evaluation tables in **Chapter 3**, CFS subset evaluator returned satisfactory results for all types of datasets, particularly for the Binary Class dataset, revealing the power of the algorithm. Comparing the features chosen by the CFS algorithm with those generated by other feature selection techniques shows that lot of the features in the CFS subset were also found on the top ranked genes from other methods. These results are consistent with CFS not picking random features, and the results obtained are similar to other feature selection techniques, as shown in **Chapter 2**. Therefore, in NetCFS algorithm were used gene sets generated by CFS algorithm.

The main steps of NetCFS algorithm are the following:

1. Gene expression datasets (microarrays) are used as input, genes as variables and samples as observations.

2. CFS is applied and a gene subset that best differentiates the classes is selected (as described in **Chapter 2**)

3. From the selected gene subset (**step 2**), CFS is re-applied using each gene as predictor and picking a new subset of genes whose expression is predictive of the expression of the target gene.

4. GENIE3 and the random forest algorithm are used then to generate an $M \times M$ matrix consisting of a ranking between the target genes and its predictors.

For the implementation of the algorithm, matlab version of GENIE3 is used as well WEKA API and java classes. To implement statistics and display networks, the Cytoscape software was used (Shannon *et al.* [2003]).

Figure 4.8 shows a schematic representation of the NetCFS algorithm to generate gene expression networks and network statistics. The gene expression matrix is converted to ARFF format to be read by WEKA application and a WEKA object is generated. This object is passed to the "Generate Matrix"

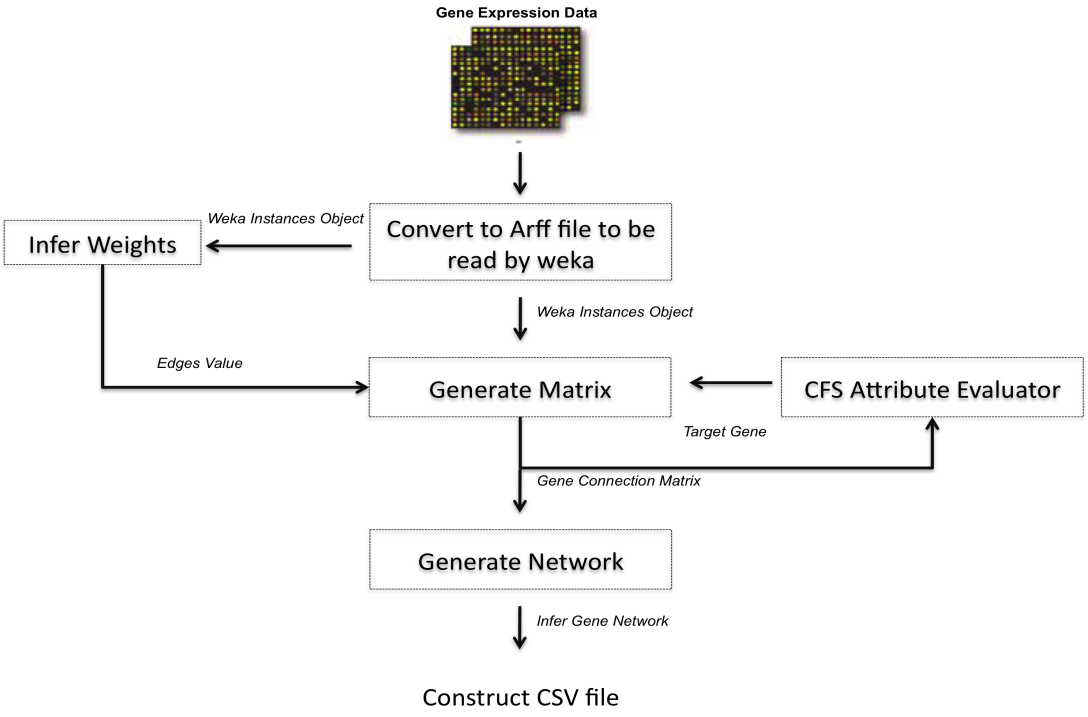

Figure 4.8: Schematic representation of NetCFS software application (more details of the functions are presented in Appendix B).

process that runs CFS for every gene in the dataset (details about the functions of "Generate Matrix", as well of the java implementation of NetCFS, are given in **Appendix B**). The result of this process is an $M \times M$ matrix where M represents the number of genes. To generate the weight of the edges, GENIE3 or random jungle and the resulting weights are used with the network generated from the CFS to produce the final network. The process "Generate Network" creates a CSV file containing edges between target node to source node and their weights. The application also contains a number of features that will be used to test the networks generated, including generation of random networks and co-expression networks derived from Pearson Correlation Coefficient.

4.3.2 Application in Leukemia Dataset

Leukemia dataset described in **Chapter 3** was used to infer networks using Pearson Correlation Coefficient (PCC) and NetCFS. Co-expression networks were constructed for both classes (ALL and AML) by applying cut-off value greater than 0.6 PCC, which is shown previously that generates biologically relevant networks (Elo *et al.* [2007], Prieto *et al.* [2008]). NetCFS was implemented in both classes to generate directed and undirected networks. To evaluate the statistical significance and stability of the networks derived from the two methods, random networks were also generated by applying 100 permutation in the expression dataset. The simulation showed that on average, nodes in the random networks exhibit low number of connections (low degree nodes) proving the statistical significance of the high connected nodes found from NetCFS and PCC.

Random networks follow a Poison distribution while networks derived from PCC and NetCFS obey to a power law distribution, indicating that these networks are scale free. These findings are supported by the statistics in Table 4.3. AML class was tested and networks derived from NetCFS (undirected) and PCC were highly structured with high connectivity (average degree range: 9.3 to 54.5) and low clustering coefficient (range: 0.03 to 0.4). Note here, that the degree distribution from NetCFS is very low compared to the networks derived from PCC, which adhere to the properties of small-world networks with only few highly-connected genes, and most genes having only a low number of interaction partners

A small-world network has a small diameter and a large clustering coefficient (see section 4.1), which is believed to be related to an efficient and controlled flow of information

Table 4.3: Topological parameters of networks derived from NetCFS and PCC

Network	Average Degree	Diameter	Clustering Coefficient
AML (undirected)	9.356	9	0.041
AML PCC 0.7	54.526	6	0.036
AML PCC 0.8	27.587	9	0.32
AML PCC 0.9	15.634	11	0.45

The networks were compared on the basis of the hub genes generated. Hubs were defined as nodes with a degree of 75 or greater and they obtained by the CFS algorithm for ALL and AML networks. These genes were compared to the same genes found in the respective Co-expression Networks (created with different cutoffs). **Table 4.4** shows the result for class AML.

Table 4.4: Hub genes according to the number of their neighbors

Gene Name	CFS	PCC 0.6	PCC 0.7	PCC 0.8	PCC 0.9
KIAA0001/P2RY14	78	423	231	136	92
Small GTP-binding protein, S10	108	127	32	4	0
IgG Fc binding protein	76	109	27	3	1
KIAA0237 gene	119	112	25	6	1
Glucocorticoid Receptor, Beta	103	274	150	97	1
Centrin mRNA	85	293	104	24	2
Adenylosuccinate synthetase mRNA	77	165	45	7	0
Rho GDP-dissociation Inhibitor 1	116	181	57	6	0
TFE3 transcription factor gene	75	212	84	15	0
TIM17 preprotein translocase	137	85	18	3	0
TNR Tenascin R	92	98	26	3	1
Spermine synthase	116	167	57	11	0
E2F5 E2F transcription factor 5, p130-binding	92	129	48	6	1
Thymopoietin (TMPO) gene	79	111	40	8	0
Polyadenylate binding protein II	92	211	75	14	0

The results presented here are preliminary for the NetCFS algorithm and focus on the hub genes generated by the two methods. As shown in Table 4.4, NetCFS exhibit core genes significant for the phenotype of interest that were not detected (as hubs) by stringent thresholds of Pearson Correlation Coefficient (PCC). Interestingly, NetCFS identified hub genes that could not be discovered with stringent PCC thresholds. For instance, targeting *IgG Fc binding protein* was found to be an alternative strategy to gene therapy on human acute myeloid leukemia (AML) blasts by enhancing their immunogenicity for autologous T cells (Notter *et al.* [2001]). Similarly, *P2RY14*, belongs to the family of G-protein coupled receptors and has a transcript variant (CB2) which has been previously found to be involved in certain cases of human AML (Jorda *et al.* [2003]). These

observations indicate the potential significance of NetCFS as a tool to identify interesting candidate genes for further study.

4.3.3 Discussion

In this section, a novel algorithm, NetCFS, is introduced and implemented as a JAVA and Matlab application to infer gene expression networks. This algorithm is based on GENIE3 where instead of feature selection using random forest a subset generator named Correlation Feature Selection (CFS) is used. Comparing to networks derived from Pearson Correlation Coefficient (PCC), it is clear that NetCFS produce weighted networks. Exploiting the classification algorithm, CFS, the algorithm reflects the underlying causal interactions among the genes since a link is displaying the influence of the target gene by the expression of another gene. Furthermore, using CFS, the algorithm refrained from the problem of thresholding.

Overall, leukemia application was a benchmark set for testing the algorithm which is still in first stages. Future research in causality and clustering of the gene networks that are generated from NetCFS should be applied.

4.3.4 Disclaimer

The experiments and data acquisition for melanoma samples were done by Dr. Karagiannis Panagiotis. Analysis of data and network construction were done by myself. I was the sole contributor for the design of NetCFS method and the algorithm was developed in collaboration with Mr Clint Mizzi

Chapter 5

Data Mining in Patient Stratification

This chapter presents an integration of supervised and unsupervised learning as a way forward to stratified medicine. Not all patients respond similarly to different treatments. Progress in the understanding of disease mechanisms and drug actions are opening opportunities to match therapies to patient populations and therefore may deliver novel approaches in personalised patient medical care.

Section 5.1, gives an overview of the potential biological relevance of using ensemble decision tree predictors to determine molecular disease subtypes, in what may initially appear to be a homogenous clinical group. With the computational pipeline presented here, we can achieve classification of transcriptional patterns into appropriate sub-groups that can lead to significant insights into disease mechanisms and novel, targeted therapeutic approaches (**sections 5.2** and **5.3**).

5.1 Disease Molecular Sub-Types Through Ensemble Decision Trees (EDTs)

Psoriasis is one of the most prevalent chronic inflammatory disorders. It is caused by an interplay of genetic factors and the environment on the background of dysregulated immune system (Nestle *et al.* [2009]). The disease affects 2 - 3

% of the population worldwide (Lebwohl [2003]) and has variable morphology, severity and distribution. There are several clinical variants of psoriasis, but the most common variant, plaque psoriasis, is characterised by chronic, symmetrical, silvery-scaled, sharply circumscribed plaques (Nestle *et al.* [2009], Lowes *et al.* [2007]). Plaque psoriasis is the most common form of the disease and can begin in childhood and late adolescence (Type 1) or in adulthood (Type 2), with a predilection for elbows, knees and the scalp.

Although the cause of psoriasis remains unknown, it is thought to be a complex and multifactorial disorder brought about by the combination of multiple susceptibility genes (Capon *et al.* [2007], Liu *et al.* [2008], Zhang *et al.* [2009]), a dysregulated immune system (Volpe *et al.* [2008]) and environmental factors (Krueger [2002]). Through Genome Wide Association Studies (GWAS) (Feng *et al.* [2009], Strange *et al.* [2010]), a number of genetic variants have been identified as contributing towards psoriasis pathogenesis. A unifying model that integrates genetic, environmental and immunological aspects of skin inflammation has been proposed (Valeyev *et al.* [2010]).

In recent years, progress has been made in understanding the pathogenesis and treatment of psoriasis. Pathogenesis is mainly linked to activation of several types of leukocytes that control cellular immunity and to a T-cell-dependent inflammatory process in skin that accelerates the growth of epidermal and vascular cells in psoriasis lesions. More details about the role of cytokines and different cell-types in psoriasis pathogenesis are given in **Chapter 6, section 6.1**. Current therapeutic approaches against the disease take advantage of proteins or antibodies aiming either at specific inflammatory co-activators or more generally at immune cells (Lowes *et al.* [2007]). While there is now increasing insight into the genes conferring disease susceptibility, much less is known about the types of regulatory networks of expressed genes which define the molecular signature of the disease.

The first large-scale and detailed gene expression studies of psoriasis identified various differentially expressed genes by comparing non-lesional and lesional skin against normal tissue (Bowcock *et al.* [2001], Oestreicher *et al.* [2001], Zhou *et al.* [2003], Haider *et al.* [2006]). Recent studies have attempted to elucidate the molecular pathways underlying in psoriasis (Gudjonsson *et al.* [2009a], Gudjons-

son *et al.* [2009b], Gudjonsson *et al.* [2010],Suarez-Farinas *et al.* [2010]). However, determining genes that contribute to complex human disorders through analysis of microarray data is challenging due to the large number of gene predictors, their possible interactions, and the small number of samples. Termed the "small n, large p" problem (Strobl *et al.* [2007]), this implies that classical statistical methods cannot be implemented directly in functional genomics approaches for the identification of diagnostic or prognostic biomarkers. In this respect, decision trees have proven to be a sensible non-parametric method for classification and variable selection (Breiman *et al.* [1984]). Random forest (RF) classification is an ensemble of CART decision trees and has been found to outperform other machine learning techniques for analysis of microarray data (Diaz-Uriarte & Alvarez de Andres [2006], McKinney *et al.* [2006], Heidema *et al.* [2006]), as discussed in Chapter 2.

In this part of the thesis, a computational methodology based on decision tree predictors is developed to discover molecular sub-groups from gene expression data and illustrate gene signatures associated with each group. The random forest (RF) algorithm (Breiman *et al.* [1984]) is used here to **(i)** cluster psoriasis transcripts into subgroups and **(ii)** discriminate between disease phenotypes by generating gene signatures that best differentiate them. RF has been shown to be robust in noisy data, to avoid over-fitting in cases where the number of features is larger than the number of observations and to be particularly suitable for the feature selection process (Diaz-Uriarte & Alvarez de Andres [2006], Jiang *et al.* [2009], Hastie *et al.* [2009]).

More specific to the current analysis, first gene expression profiles in normal and disease skin tissues were analysed, so as to define common differentially expressed genes. This core gene set was then used to group psoriatic tissue samples through RF clustering of real and synthetic data. This step resulted in dividing psoriatic tissues into two subgroups according to similarity of gene expression patterns. Finally, RF classification was used to derive gene signatures able to discriminate between normal and disease phenotypes, including the above-proposed new psoriatic subgroups. Such gene signatures are discussed in following sections with respect to their effect on defining distinct molecular characteristics and were validated through comparisons with other psoriasis gene expression

studies.

Molecular profiling of psoriatic phenotypes followed by classification of tissue samples into appropriate disease classes has the potential to derive clusters of similar transcription responses from the entire repertoire of profiles generated. Especially in the case of a homogeneous clinical patient group, such as plaque-type psoriasis, the classification of transcriptional patterns into appropriate subgroups may reveal distinct molecular mechanisms that may operate within this group and may explain variability in response and options of disease treatment. Overall, given the predictive nature of the decision model employed, such patient categorisations can lead to significant insights into disease mechanisms and novel, targeted therapeutic approaches.

5.2 Materials and Methods

5.2.1 Data sources

Microarray data on psoriatic gene expression were obtained from the Genetic Association Information Network (GAIN) Database (Nair *et al.* [2009]), available through the NCBI database of Genotypes and Phenotypes (dbGaP). These experiments describe tissue samples from 71 individuals, of which 34 were healthy controls (*NN*) and 37 patients affected by chronic plaque psoriasis. Paired samples from lesional (*PP*) and non-lesional (*PN*) tissues were extracted and gene expression was measured by microarray experiments on the Affymetrix HU133 Plus 2.0 platform. Raw data were normalized using quantile normalization and expression estimates were computed using the Robust Multichip Average (RMA) method (Irizarry *et al.* [2003]).

Analyses performed on the above dataset were validated through comparison with gene expression datasets GSE14905 and GSE13355 from the ArrayExpress database (Parkinson *et al.* [2011]). The first study consisted of 21 biopsies from healthy donors and 26 paired non-lesional and lesional plaque type psoriatic patients (Yao *et al.* [2008]) and the second dataset comprised 64 normal samples and 58 psoriatic tissues (Gudjonsson *et al.* [2010]). Both studies were conducted on HGU133plus2 Affymetrix chips.

5.2.2 Differential Expression Analysis

In order to define a core dataset of differentially expressed genes in the psoriatic phenotypes examined, pairwise comparisons between 34 normal (*NN*) and 37 lesional (*PP*) and non-lesional (*PN*) gene expression vectors were performed. The differential expression between pairs of samples (PP vs. NN, PN vs. NN, PP vs. PN) was assessed using GenePattern (Reich *et al.* [2006]). Significance scores were assigned to each probe (p-value <0.05), multiple hypothesis testing was applied with FDR<0.05 to reduce the false positives and the top ranked 5000 probes were extracted for each pair of samples. Of those, the set with the most common expression alteration among the three pairwise comparisons was selected. Probes that mapped to the same gene were averaged and the average intensity across all corresponding genes was used. A core set of 228 probes common to all three pairwise comparisons was established. Of these, a total number of 206 unique known genes were derived yielding 130 up-regulated and 76 down-regulated genes.

Hierarchical clustering and principal component analysis (PCA) were implemented to identify distinct patterns of gene expression within the "core" 206 differentially expressed genes. The PCA procedure was implemented as part of the PCA package in R[1]. Hierarchical clustering heat-maps were generated in R based on Euclidean distance. $Z - scores$ were calculated from the level of normalized expressions of 206 genes according to the mean and standard deviation of a reference set (control samples, NN).

PCA and Hierarchical clustering were used to confirm the separation of the three skin-types and then Random Forest was used in an unsupervised way to identify sub-types/sub-groups within the same skin type, e.g. psoriasis cases.

5.2.3 Decision Tree Classification Model

An ensemble of decision trees model was built according to the random forest (RF) classifier using a deterministic algorithm (Classification and Regression Tree Algorithm, CART) (Breiman *et al.* [1984]). Given a gene expression matrix, a RF classifier was constructed to classify tissue samples into relevant disease classes (NN, PN, PP) based on gene expression measurements (variables). Details on

[1] www.r-project.org

the classification strategy of RF are given in Chapter 2. Variable importance measures were implemented through mean decrease in accuracy and the gini index (GI) (Diaz-Uriarte & Alvarez de Andres [2006]), to find the genes that best discriminate between the different disease phenotypes. Both measures were tested and were found to correlate well. The gini index was therefore adopted to express the relative effects of gene predictors in determining the relevant disease classes (see **Chapter 2**). To estimate the empirical p-value for GI, 1000 permutations of the tissue samples were implemented and the importance values were recalculated for the permuted data set. The maximum Gini Index over all the genes in every permutation was recorded and thereby an empirical distribution of the maximum importance was estimated, as in similar analyses (McDonough *et al.* [2009], Sohn *et al.* [2009]).

5.2.4 Clusters of Disease Sample sub-groups through Ensemble Of Decision Trees (EDTs) classification

A procedure to generate sub-groups of disease samples from gene expression measurements through the use of RF is described here. The random forest proximity measure, defined through the number of times each tree detects these samples in the same terminal node, is used as a means to express the similarity between psoriatic samples from gene expression observations (see **Chapter 2**). Synthetic data are generated by randomly sampling from the empirical marginal distributions of variables. RF classification is applied to distinguish the 37 psoriatic samples from the synthetic data and the dissimilarity matrix is used to indicate distances between psoriatic samples, as previously (Shi *et al.* [2005]). Through multi-dimensional scaling, samples are represented as points before clustering through CLARA (Kaufman & Rousseeuw [1990]). This procedure was implemented in R. The sub-groups were further tested to assess their differences for various clinical outcomes (soriasis Area and Severity Index (PASI), Body Mass index (BMI), Age of Onset, Age and Body Surface Area (BSA)). The statistical significance of disease clusters with respect to clinical variables was done through Wilcoxon signed-rank test.

5.2.5 Methods applied for Network and Functional Enrichment Analysis

As discussed in **Chapter 4**, network representation of a disease phenotype can give a better insight of the molecular mechanism of the disease. Pairwise Pearsons correlation coefficient is estimated for the 206 differentially expressed genes that were common in all tissues. A similarity matrix was calculated for each skin sub-type and a co-expression network was visualised using the Cytoscape software. Markov Cluster Algorithm (MCL) was used to generate the interacting groups (clusters) via genes sharing higher-order connectivity in their local neighborhoods (Enright *et al.* [2002]). To assess statistically significant enriched pathways involved in the four different skin groups, p-values were calculated using the hypergeometric statistical test and False Discovery Rate (FDR<0.05) was used to correct for multiple comparisons. The default background distribution is considered to be the whole genome. Pathway enrichment analysis was performed using the ReactomePA package in Bioconductor (Yu *et al.* [2012]).

5.3 Pipeline for patient stratification

A pipeline (**Figure 5.1**) for patient stratification according to gene expression profiles in psoriasis was implemented to generate molecular sub-groups and uncover gene signatures associated with each disease group. Such an approach has possible predictive implications: given relevant expression measurements for key signature genes, uncharacterised tissue samples can be ascribed to these predefined disease classes, which can reflect different disease prognosis or response to treatment.

5.3.1 Gene expression patterns define a core set of dysregulated genes among normal, non-lesional and lesional skin

Skin samples from psoriatic patients were either of inflamed, lesional type (PP, involved skin) or non-inflamed, non-lesional tissue (PN, uninvolved skin). These

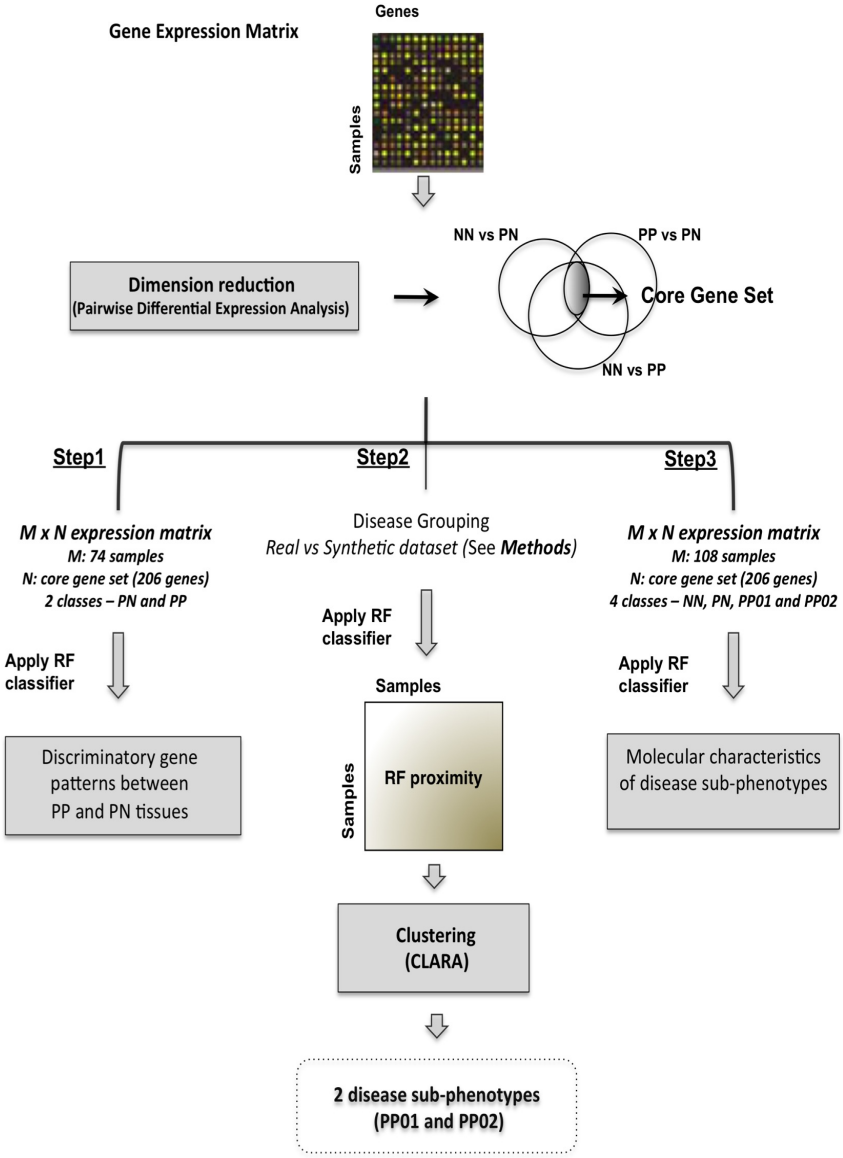

Figure 5.1: Pipeline for patient stratification

were analysed together with skin samples from normal individuals (denoted as NN). Differential expression analysis was performed and involved three pairwise comparisons between tissue datasets (i.e. NN vs. PN, PN vs. PP and NN vs. PP), resulting in three sets of differentially expressed probe sets per pair (p-value$<$ 0.05 and FDR$<$0.05). A set of 228 probes was shared across all datasets (**Figures 5.1** and **5.2a**) that corresponded to a total number of 206 unique genes, of which 130 genes were over-expressed and 76 under-expressed in PP samples compared to NN. This group of genes constituted a core set of genes expressed differentially across the three disease phenotypes and was used to derive disease-specific expression patterns in the RF-based procedure described in the following sections.

Unsupervised hierarchical clustering was carried out on the set of 206 core genes to explore and visualise the patterns of gene expression from normal (NN) to non-lesional (PN) and then to lesional (PP) skin samples. **Figure 5.2b** shows an overview of gene expression for the core probe sets, clustered according to similarity of expression across NN, PN and PP samples. This visualisation represents a striking outline of the varying transcriptional patterns at each disease phase, progressing gradually from generally non-differentiated gene expression in non-inflamed tissues (NN, PN), to markedly differentiated genes in lesional samples (PP).

Principal component analysis (PCA) was used to assess the clustering of samples when progressing from un-inflamed to inflamed skin. There was a clear distinction between lesional (PP) and non-lesional (NN and PN) phenotypes (**Figure 5.2c**), manifested as distinct clusters of samples from normal to the involved phenotype through non-involved skin. Normal and psoriatic un-involved samples (NN and PN) co-clustered away from involved cases (PP), in agreement with previously published analyses (Gudjonsson *et al.* [2009b], Zhou *et al.* [2003]). This demonstrated the changes in gene expression profiles across NN, PN and PP skin and revealed a marked difference between inflamed (PP) skin and un-inflamed (PN and NN) phenotypes.

Among the strongly dysregulated genes in the core gene set, several of the under-expressed genes were found to encode proteins involved in fibrotic processes and immune responses. For example, *FN1, PDGFC, MYH10* are involved in the

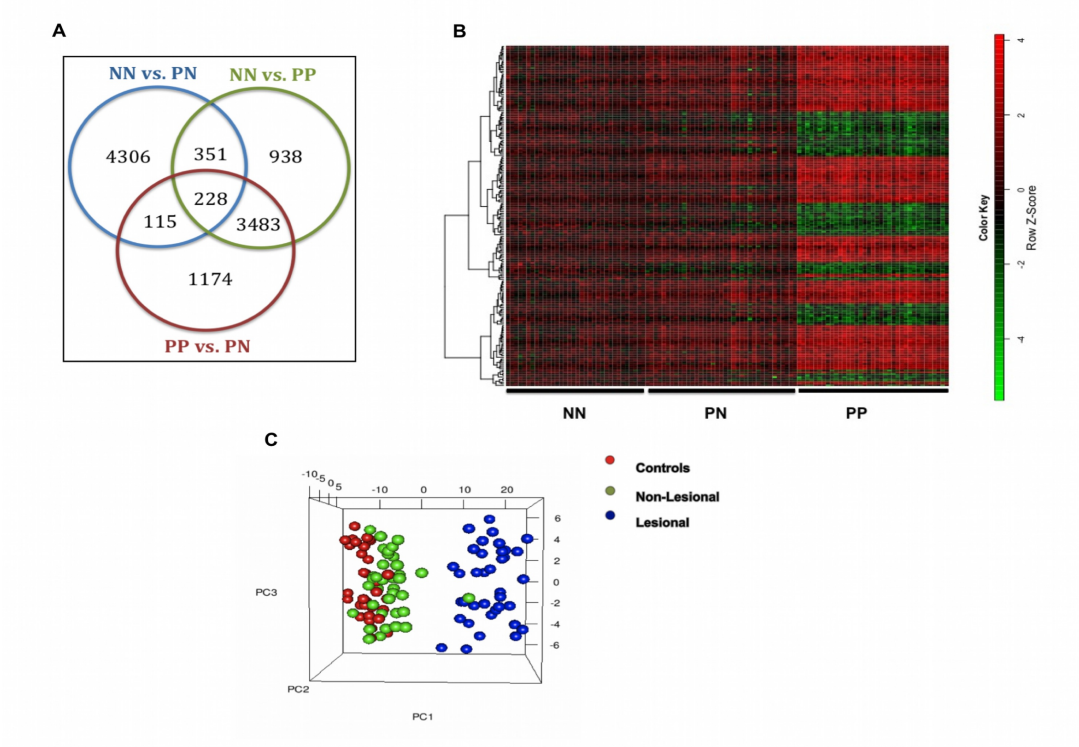

Figure 5.2: (A) Gene expression was analysed to reveal probe sets that were differentially expressed between pairwise comparisons of PP, PN and NN tissue groups. The Venn diagram shows the number of probe sets identified in each of the three differential analyses performed. Probe sets common to all three pairwise comparisons were 228 (206 genes). (B) Microarray analysis of 108 skin tissue samples (in columns) for 206 genes (in rows) common to all tissue types, identified through differential expression analysis. Tissues have been grouped according to disease phenotype (normal NN, non-lesional PN and lesional PP) and heatmap colours indicate z-score of each gene expression value against the mean of corresponding normal values, (green: decreased expression, red: increased expression, inset). Similarity of gene expression vectors across all samples is represented by the dendrogram on the left. (C) Principal Component Analysis to suggest sample clustering across skin types according to gene expression patterns. Good separation of inflamed (PP) and non-inflamed (PN, NN) tissues was observed, indicating a progression from normal (red) to lesional skin (blue) through the non-lesional cases (green)

regulation of the actin cytoskeleton, which participates in fundamental processes such as the regulation of cell shape, motility and adhesion [30]. *DIXDC1, CGNL1* and *SSPN* encode cell adhesion and junction proteins. Betacellulin (*BTC*), *IL1F7, CD81, FN1, PDGFC* and *SCARB2* are immune response genes. In addition, *MEGF9, BTC, FN1, PHF2* belong to the family of growth factors that activate the epidermal growth factor receptor,*EGFR (ErbB1)* and according to a previous study *BTC* plays an important role in skin morphogenesis (Schneider *et al.* [2008]). Among the over-expressed genes, several participate in keratinocyte proliferation and differentiation (*EREG, KLK8* and *PPARD*). Of note is *KLK8*, potentially involved in the modulation of hyperkeratosis in a psoriatic lesion and may be implicated in preventing excessive keratinocyte proliferation, resulting in increased shedding of corneocytes. This is clinically reflected in the copious quantities of scale that are shed by psoriasis patients (Kishibe *et al.* [2007]). Genes *LTB4R2* and *PPARD* are also involved in keratinocyte migration. Finally, a group of up-regulated genes *SNRPG, SNRPD1, SNRPD3, SNRPA1, SNRPC, SF3B14, SFRS9* is involved in spliceosomal assembly. Overall, most dys-regulated genes were found to be consistent with current knowledge.

5.3.2 Distinctive gene expression patterns between lesional and non-lesional tissues (PP vs. PN)

Following the general patterns of psoriatic tissue differentiation, the use of decision tree ensembles was explored to classify samples into PN and PP classes and derive the major gene patterns able to discriminate the psoriatic phenotypes (see **Figure 5.1**, **step1**). We used 74 tissue samples from psoriasis patients, each characterised by a vector of core gene expression values, and a random forest (RF) classifier (Breiman *et al.* [1984]) was applied to distinguish samples in lesional (*PP*) and non-lesional (*PN*) phenotypes. The classifier employed 1000 trees with training of each tree performed on $\frac{2}{3}$ of samples and testing on the remaining $\frac{1}{3}$. The prediction accuracy of the classifier was high (accuracy 97.3 %, OOB error rate 2.7 %).

The random forest classifier was then used to indicate the relevant importance of features in the classification, i.e. which genes are more important in predicting

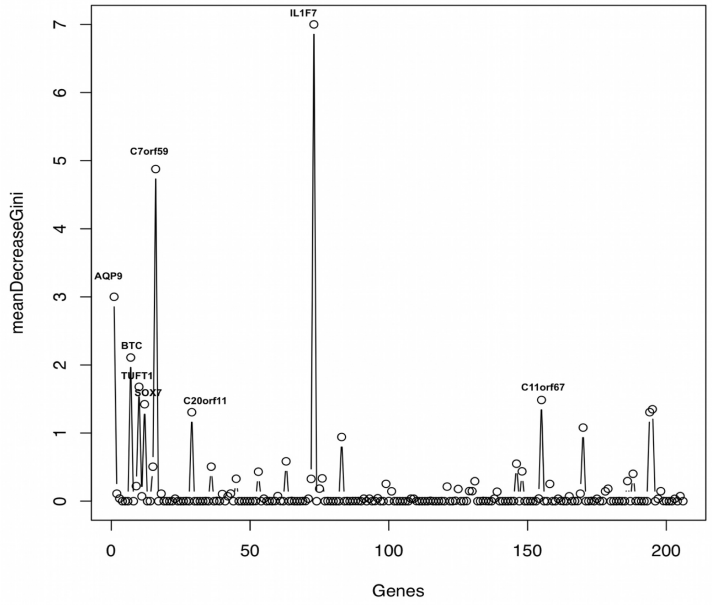

Figure 5.3: Informative genes for the classification of skin samples in lesional and non-lesional classes (PP and PN, respectively). Gini index (GI) was used to generate a variable importance measure and provide an estimate of feature (gene) relevance to disease state. The five most important genes in determining disease classes were IL1F7, C7orf59, AQP9, BTC and TUFT1

the appropriate disease class. Genes were ranked through the gini index (GI) in terms of their discriminative power (see Methods) and **Figure 5.3** shows genes with the highest GI when distinguishing inflamed (PP) from non-inflamed (PN) skin. Five genes indicated through this procedure were *IL1F7, C7orf59, AQP9, BTC* and *TUFT1* and were all related to immune response processes.

5.3.3 Identification of molecular sub-types within psoriatic tissue samples

In addition to key patterns that defined disease outcome in psoriatic tissues above, we used random forest in unsupervised mode, as a clustering platform to group

lesional psoriasis samples based on their gene expression properties (see **Figure 5.1**, **step 2**). The aim was to generate two sub-groups among disease tissues (PP), before further classification runs could identify molecular differences among them (**Figure 5.1**, **step 3**, discussed later). First synthetic data were generated by randomly sampling the gene expression observations. Then, a random forest predictor was built to distinguish the real from synthetic data and define a similarity measure between the psoriatic cases in the form of the random forest proximity measure. Finally, CLARA clustering of the proximity matrix partitioned the psoriatic cases into two groups, named PP01 and PP02 (**Figure 5.1**, **step 2**). The adjusted rand index to indicate the difference between the two identified sub-groups was -0.0269.

The relationship between these sub-groups and clinically measurable parameters, were assessed. Psoriasis Area and Severity Index (PASI), Body Mass index (BMI), Age of Onset, Age and Body Surface Area (BSA) were evaluated against subgroups PP01 and PP02. Of these, age was found to be significantly altered between the two subgroups (p-value 0.0184, Wilcoxon signed-rank test). It is emphasised here that plaque-type psoriasis constitutes a homogeneous clinical group, distinct from other forms of psoriasis. Therefore, it is not surprising that such coarse-grained clinical parameters can not capture the subtle differences in gene expression profiles in plaque psoriasis sub-groups (PP01, PP02). Instead, the focus here is to distinguish the underlying biological mechanisms, in terms of distinct biochemical pathways and interactions that act in these subgroups, as we report in following sections.

Having separated psoriatic samples into two sub-groups, a new gene expression matrix, where PP samples were split into PP01 and PP02, was used as input to another round of RF classification (see **Figure 5.1**, **step 3**). The core genes (total of 206) were used as variables to classify 108 samples in any of the four classes (normal NN, non-lesional PN, first lesional group PP01, or second lesional group PP02). The purpose of this series of experiments was to assess the discriminatory power of different genes in deriving the four disease classes through classification. The classifier showed good prediction accuracy (79.6 %, OOB error rate 20.37 %, 1000 trees). **Figure 5.4** shows the MDS plot for this classification experiment, illustrating the relative clustering of samples in four skin phenotypes. As before,

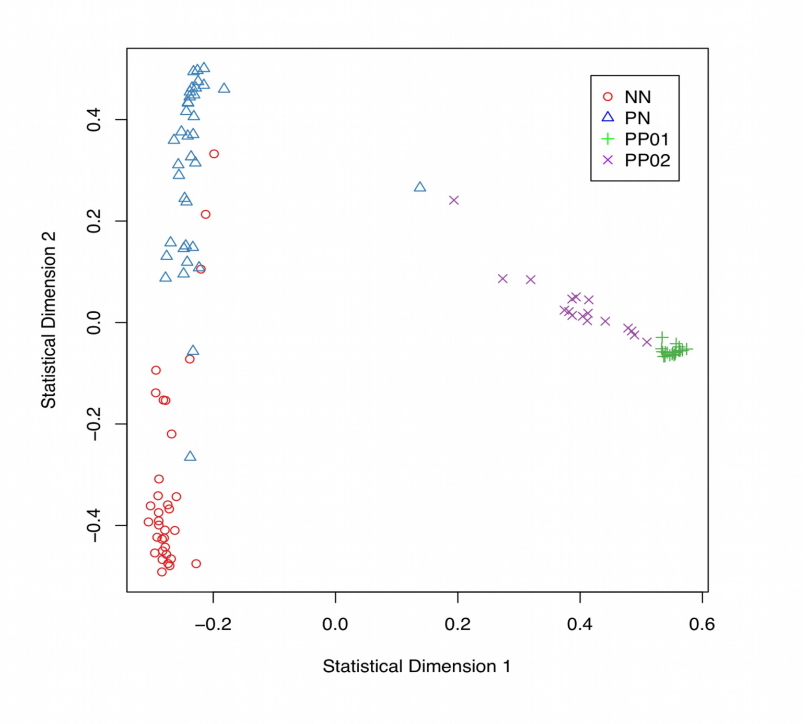

Figure 5.4: A multidimensional scaling (MDS) plot to illustrate the molecular grouping of samples. A dissimilarity matrix in random forest is constructed through the use of synthetic data drawn from the distribution of psoriatic samples (see Methods). Patients are clustered according to these dissimilarities and two distinct psoriatic groups are identified, PP01 (green) and PP02 (purple). All lesional samples (PP) cluster away from both normal (NN) and non-lesional (PN) tissue samples, in accordance to observations in **figure 5.2c**

non-inflamed tissues (NN and PN) clustered away from the inflamed tissues (PP). Additionally, the relative segregation of the two PP subgroups was also apparent.

The contribution of particular genes in differentiating the corresponding disease phenotypes was also assessed through gini index as variable importance measure and **Figure 5.5** illustrates a measure of the predictive power of genes in classification. This set of most informative genes in **Figure 5.5** were also confirmed by calculating the empirical p-value by permutating the tissue labels Wang *et al.* [2010].

To extract the differences in gene expression between the two psoriasis sub-

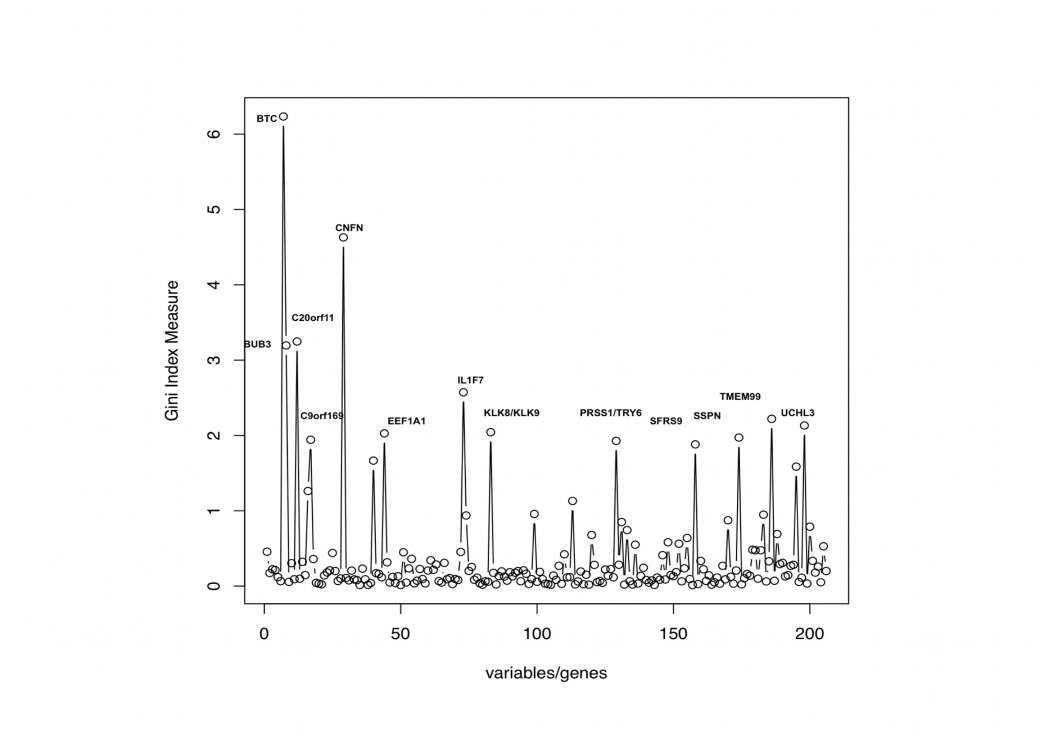

Figure 5.5: Genes identified as most informative through RF classification of skin tissues in four molecular groups (NN, PN, PP01 and PP02). Gini index (GI) was used as variable importance measure for estimating the discriminative power of relevant features (genes) and, consequently, their relevance to disease state. The five most important genes in determining disease classes were *BTC, CNFN, C20orf11, BUB3* and *IL1F7*

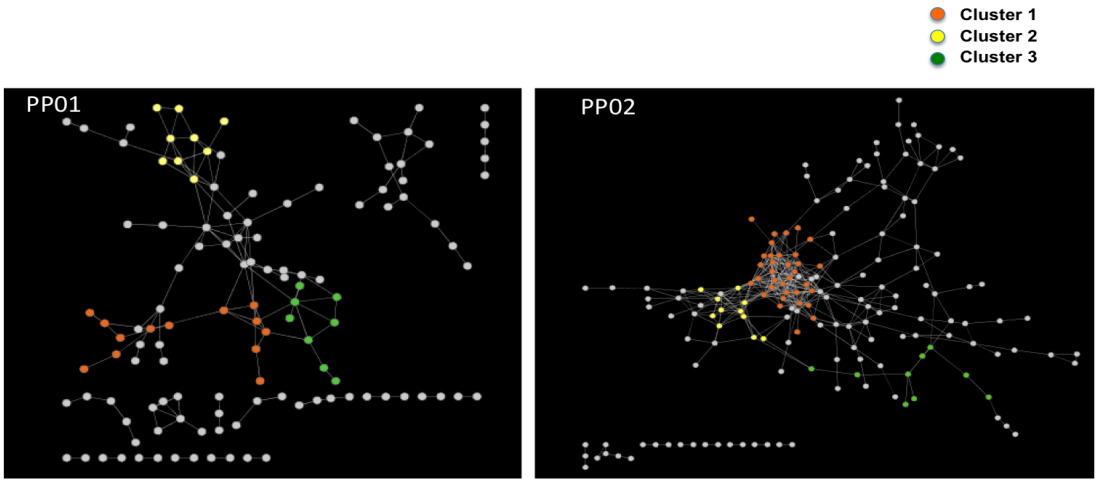

Figure 5.6: Markov Cluster Algorithm (MCL) applied on the psoriatic sub-group tissue sample networks to extract clusters of gene expression. Both networks consisted of 36 clusters and the largest clusters (number of nodes>8) for both networks are shown and denoted by colour. Pathway enrichment for these clusters is shown in Tables 6.1 and 6.2 for PP01 and PP02 networks respectively.

groups, a co-expression network was generated of the core genes for each of these groups (**Figure 5.6**). This resulted in two networks for PP01 and PP02 with different topological properties. The PP01 network consisted of 122 nodes with 142 edges, whilst PP02 had 173 nodes with 472 edges. After clustering with MCL, 36 clusters were identified for each patient group, PP01 and PP02, and functional enrichment analysis of the largest clusters has indicated different biochemical pathways linked to each of these groups. The PP01 network clusters were enriched in signalling pathways, such as Wnt Signaling pathway, Notch, TGF-beta, ErbB and mTOR signalling pathways, whereas clusters in PP02 network were more involved in metabolic pathways (**Tables 1** and **2**). This indicated that the two lesional psoriatic sub-groups possess different functional properties, suggesting different underlying biological processes.

5.3.4 Key genes associated with disease sub-classes and comparison with other studies

Variable importance analysis was used to derive highly discriminative genes in classification of disease sub-types. As described previously, the gini index (GI), calculated from RF classification (108 samples, 206 genes, 4 classes, see **Figure 5.1**, **Step 3**), was used to rank each of the 206 genes in terms of their discriminatory effect in assigning samples in each of the four tissue groups (NN, PN, PP01, PP02). The 43 genes with highest GI were schematically represented in a circular layout (**Figure 5.7**) to show their effect in each of the four classes. For example, *BTC*, part of the ErbB and ERK Signaling pathways, has GI equal to 0.11 for samples classified in the PP01 class and a significantly lower GI when classifying samples in the other tissue groups (GI=0.041 for NN, GI=0.056 for PN and GI=0.055 for PP02). Similarly, the importance of the other 43 genes in classifying samples in the four phenotype classes was determined and is illustrated (**Figure 5.7**).

In the NN group, *CNFN* and *BUB3* were more frequently selected to define a split in the classification trees of the forest, whereas *BTC, IL1F7* and *TMEM99* (GI>0.02) were important in classification of PN samples. Within the PP group, *BTC, C20orf11, EEF1A1, CNFN, IL1F7, PRSS1/TRY6, SSPN* and *UCHL3* showed high discriminative value for identifying the PP01 sub-group, whereas *BTC, CNFN, IL1F7, KLK8/KLK9* and *TXNDC4* exhibited high importance for the PP02 sub-group. To further support linking these genes to psoriasis-related biological mechanisms, the PubMatrix tool was used to look up the discriminatory genes in the context of eight terms, includingpsoriasis, NK cells, T cells, immune response, Wnt signalling pathway, Notch signalling pathway, TGF beta signalling pathway and ErbB signalling pathway [35]. Out of 43 genes, 24 genes were found to occur together with these terms in biomedical literature, as seen in **Figure 5.8**. Interestingly, *BTC*, which exhibited a high discriminative value when characterising PP01 sub-type, was found to be related with ErbB signalling pathway. The latter was a highly enriched pathway in this sub-group and indicates a potential therapeutic target. *IL1RN* also had a high GI for samples classified as PP01 and was previously found to be highly related with T cell activation and

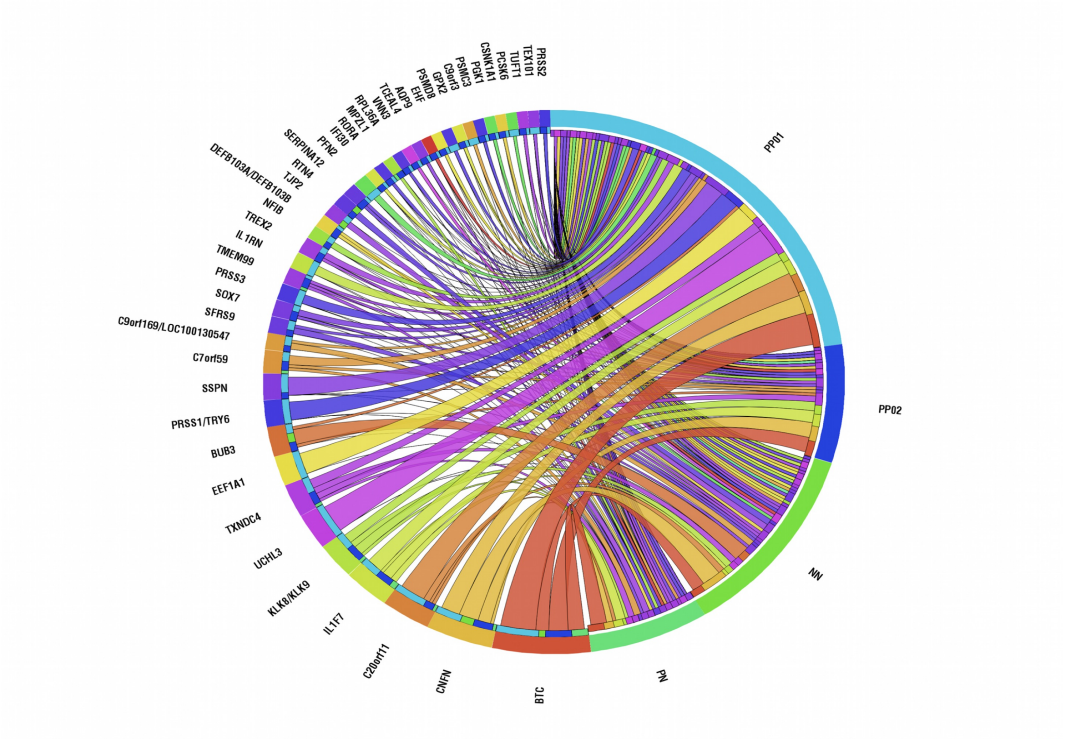

Figure 5.7: Graphical representation to illustrate the relation between 43 highly discriminative genes and disease sub-groups. Contributions shown according to gini index, calculated from random forest classification. The four skin-types (PP01: light blue, NN: green, PP02: blue, PN: light green) followed by relevant genes are arranged clockwise. Skin groups and genes are ordered according to shared pairing links

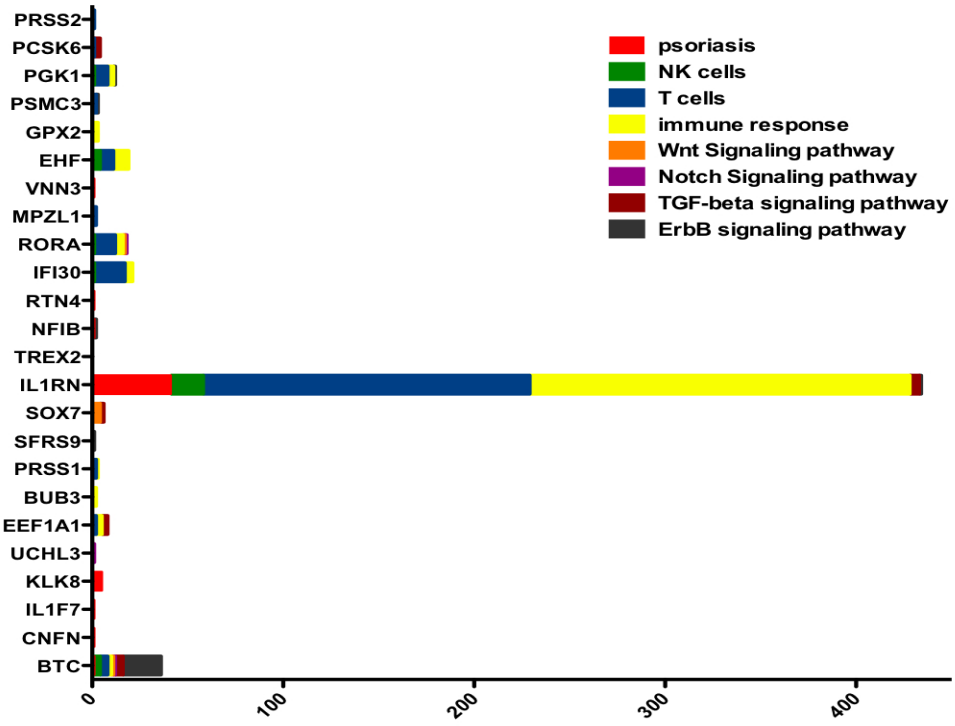

Figure 5.8: Text mining results for validation process according to the literature. Co-occurrence of gene names with disease-related terms, such aspsoriasis, NK cells, T cells, immune response, Wnt signalling pathway, Notch signalling pathway, TGF beta signalling pathway and ErbB signalling pathway was searched in Pubmed abstracts through PubMatrix

immune response.

The pipeline outlined above was replicated with two other psoriatic datasets from (Gudjonsson *et al.* [2009b])(Gudjonsson dataset) and (Yao *et al.* [2008]) (Yao dataset). Skin samples were grouped into sub-types according to their gene expression patterns, as for the GAIN dataset, using similarities derived from the proximity matrix through random forest algorithm (**Figure 5.9**). By comparing across the three datasets and the relevant gene signatures derived, the importance of specific genes was noted. *BTC, CNFN, IL1F7* were important discriminant genes in the GAIN data, while *SNRPC* and *SMURF2* played a greater role in the Yao and Gudjonsson datasets. Generating a consistent outcome of gene signa-

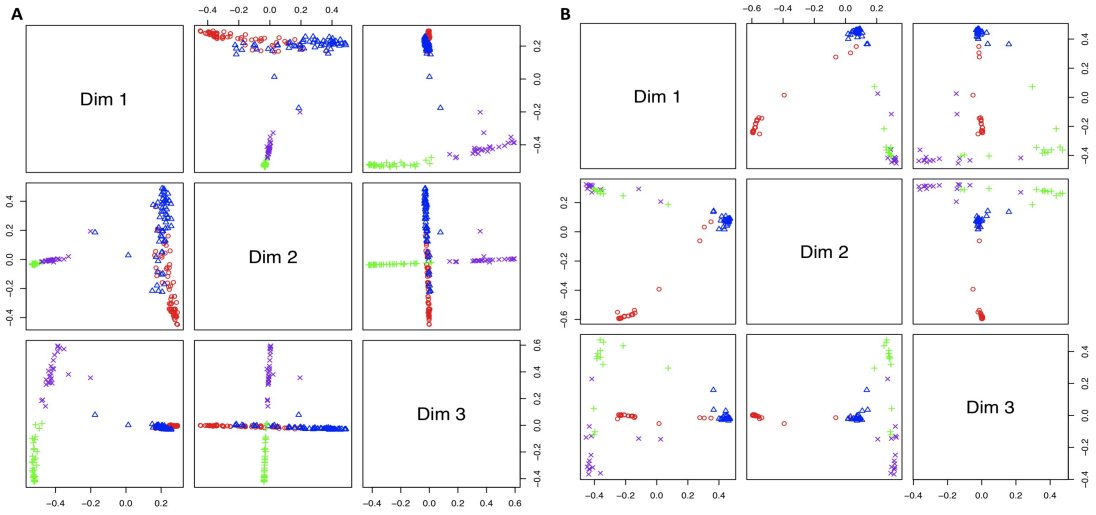

Figure 5.9: A multidimensional scaling plot of psoriasis datasets from (A) Gudjonnson et al. 2010 and (B) Yao et al. 2008 to illustrate grouping of samples according to random forest clustering.

tures across all datasets is challenging, as patient cohorts may vary significantly. Although the Yao data seem difficult to reproduce, considerable similarity exists between the other two datasets. Specifically, one of the disease subgroups in these dataset points to pathways related to Notch signalling, ErbB and TGF beta suggesting that this group may be more amenable to related therapeutic options.

Note that evaluation of psoriasis transcriptome has been assessed elsewhere (Suarez-Farinas *et al.* [2010]) and the observed low reproducibility across various studies was attributed to wide variability in clinical protocols, platforms and sample handling among different datasets. It is envisaged that the application of the present and similar strategies for predictive modelling and stratification of expression patterns, as well as the availability of larger patient studies will

bridge the disparities between various studies and yield a sharper picture of gene contributions to this complex disorder.

5.4 Discussion

Large-scale genome characterisations, through the analysis of gene sequence and expression data, are gaining increasing interest and have the potential to greatly improve our understanding of the physiological and molecular mechanisms underlying disease pathogenesis and progression. Various models of data stratification and identification of patient groups through various data mining protocols are used to support a decision making process in biomedicine. Here the use of random forest to partition psoriatic tissues in appropriate disease groups is illustrated and estimates of relevant gene predictors were generated.

Psoriasis is a common, complex immuno-genetic inflammatory disease of primarily the skin. The underlying genetics of the disease are complex with numerous implicated susceptibility genes, where replication of single loci has been confirmed for only a handful of these genes. Patients suffering from psoriasis can exhibit a host of different clinical phenotypes and response to therapy is varied and unpredictable, even within a similar clinical phenotype, suggesting underlying transcriptional differences between and within the clinical groups. The ability to investigate the underlying immuno-genomic components of these clinical sub-phenotypes has not been a possibility, until now. Identification of different transcriptional signatures and their associated molecular pathways contribute toward defining a set of biomarkers, which could serve as diagnostic and therapeutic responder tools. A computational strategy is outlined to identify molecular sub-types and corresponding putative biomarkers that may be crucial in the understanding and prediction of disease pathogenesis. Of the 206 common differentially expressed genes identified between normal, psoriatic lesional and psoriatic non-lesional groups, 130 genes (63.1%) were up-regulated and 76 genes (36.9%) were down-regulated. Dysregulated genes discovered in this study were involved in epidermal cell modulation, cell cycling and immune responses.

Microarray analysis of gene expression has been widely used to differentiate lesional and non-lesional skin of psoriatic patients (Nomura *et al.* [2003], Lee

et al. [2004]). Recently, large-scale analysis using whole genome array platforms on numerous patients per sample group have been undertaken with the aim of identifying gene expression profiles associated with a specific psoriatic phenotype (Liu *et al.* [2008], Zhang *et al.* [2009], Feng *et al.* [2009], Swindell *et al.* [2012]). In this chapter, a method is presented for identifying sub-phenotypes of lesional skin from psoriasis patients based on patterns of gene expression that characterise each group and differ significantly from normal human skin. This approach is based on a decision tree analysis of gene expression data, the extraction of associations among gene expression patterns and the identification of functional annotations and molecular signatures.

The random forest decision tree model was applied to lesional skin group to derive patient sub-groups (PP01 and PP02), which are characterised by specific differentially expressed genes. The PP01 group was defined by the up-regulation of HLA-E, which is the inhibitory ligand for innate NK cells. HLA-E takes part in processing and presenting antigen to innate immune cells. The PP02 group had more up-regulated genes related to the cells of the adaptive immune system such as CTLA-4 (associated with modulation of T helper responses), IFI30 (involved in MHC Class II antigen processing), IL4IL (immunomodulatory enzyme produced by dendritic cells), PTPN2 (associated with autoimmune disorders such as type 1 diabetes mellitus and Crohns disease) and most interestingly SERPINB8, which has been identified through Genome-Wide Association Studies (GWAS) as a new psoriasis susceptibility locus in the Chinese population (Sun *et al.* [2010]).

With regards to mechanistic details of the pathways that operate in psoriatic sub-groups, the ErbB signalling pathway has been identified for subgroup PP01 (Table 6.1). This pathway consists of a family of four related receptor tyrosine kinases (ErbB1-4) which, when activated trigger many different signal transduction pathways leading to increased proliferation, survival, motility, and invasiveness. All of these responses are important aspects of wound healing and psoriasis has many elements in common with wound healing. The main clinical feature of psoriasis relates to the thickened epidermis as a result of what may initially have been an epidermal barrier insult. An attempt to restore epidermal integrity is reflected in the activation of the ErbB signalling pathway. However in psoriasis it is possible that this pathway, along with other signalling pathways

is dysregulated (Citri & Yarden [2006]).

Other signalling pathways seem to be in effect in psoriasis sub-group PP02 (Table 6.2), for example signalling by BMP. Bone morphogenetic proteins (BMP) are members of the transforming growth factor-beta (TGF beta) superfamily and regulate a large variety of biological responses in different cells and tissues. It has been reported that BMPs are implicated in a variety of pathobiologic processes in skin, including wound healing, psoriasis, and carcinogenesis (Botchkarev [2003]).

In our analysis, when several patient clinical variables were compared across the two classes (PP01 and PP02), age was found to be significantly altered in these subgroups, indicating that this is an important factor in disease manifestations. It is worth noting that although the differences in PP01 and PP02 groups are quite marked on a transcriptional level, yet they are clinically difficult to distinguish. This observation may help explain why some patients have a different disease course to others and why some respond better to therapy than others within a given clinical phenotype. The ability to generate molecular sub-types provides putative biomarkers, which with further refinement and replication, could prove to be useful in predicting disease severity, progression and response to therapy in an individualised manner.

Random forest has become a popular Ensemble of Decision Tree (EDT) tool for analysing high-throughput genomic data. Due to the large number of variables associated with characterisation of clinical samples through gene expression measurements, reduction of dimensionality through feature selection or prioritisation is critical in disease property prediction. Through this computational analysis, there is an emerging picture of important gene predictors in psoriasis, as well as differentiation of disease in different subgroups. Future work based on richer datasets that profile larger patient cohorts, with stringent clinical phenotyping, will have the potential to draw clearer conclusions about this complex autoimmune skin disease.

5.5 Disclaimer

I am the sole contributor for the development of the computational pipeline described in this chapter.

Table 5.1: Pathway enrichment in the PP01 psoriatic group

Pathway Name	p-value	q-value
Cluster 1		
NOTCH1 Intracellular Domain Regulates Transcription	0.0008	0.0227
Signaling by NOTCH1	0.0017	0.0254
Signaling by NOTCH	0.0028	0.0275
NOTCH1 Intracellular Domain Regulates Transcription	0.0007	0.0227
Cluster 2		
Urea cycle	0.0066	0.0187
Synthesis of very long-chain fatty acyl-CoAs	0.0104	0.0187
Fatty Acyl-CoA Biosynthesis	0.0133	0.0187
Triglyceride Biosynthesis	0.0271	0.0286
Cluster 3		
PI3K events in ERBB4 signaling	0.0003	0.0054
PI3K events in ERBB2 signaling	0.0004	0.0054
Signaling by ERBB4	0.0019	0.0142
Signaling by ERBB2	0.0024	0.0142
AKT phosphorylates targets in the nucleus	0.0059	0.0289
Signaling by TGF beta	0.0106	0.0400
SHC1 events in ERBB4 signaling	0.0139	0.0400
GRB2 events in ERBB2 signaling	0.0152	0.0400
Signaling by BMP	0.0159	0.0400
SHC1 events in ERBB2 signaling	0.0165	0.0400
PIP3 activates AKT signaling	0.0205	0.0434
Immune System	0.0215	0.0434
PI3K/AKT activation	0.0263	0.0447
Nuclear signaling by ERBB4	0.0273	0.0447
GAB1 signalosome	0.0283	0.0447
Interleukin-1 signaling	0.0296	0.0447

Table 5.2: Pathway enrichment in the PP02 psoriatic group

Pathway Name	p-value	q-value
Cluster 2		
Transport of Glycerol from Adipocytes to the Liver by Aquaporins	0.0018	0.0158
Transport by Aquaporins	0.0102	0.0419
Signaling by TGF beta	0.0149	0.0419
Signaling by BMP	0.0223	0.0470
Cluster 3		
Respiratory electron transport	0.0012	0.0009
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins	0.0018	0.0009
The citric acid (TCA) cycle and respiratory electron transport	0.0036	0.0013

Chapter 6

Systems Biology For Skin Inflammation

Systems biology aims to understand the complex biological systems. The complexity of human skin due to the myriad of roles and players that are activated in different phenotypes increases the need to unravel the genomic, trascriptomic and proteomic information in the attempt to intelligently translate experimental data into clinically meaningful parameters and bring the bench closer to the bedside.

In this part of the thesis, topological and functional properties of psoriasis are presented in association with the deconvolution of cytokine-related pathways(**section 6.2**). Together, a novel systems biology approach is proposed, combining clinically relevant whole-tissue models with integrative systems analysis to identify unexpected mechanisms of disease progression and development. This integrative approach was applied to deconvolute novel IL-22-related gene sets in psoriatic disease pathogenesis and therapeutic response (**section 6.3**).

6.1 Psoriasis

In the previous chapter, psoriasis is introduced as a homogeneous clinical phenotype and report a computational pipeline for the differentiation of disease in different subgroups. Here, the environmental and genetic factors that play a role in the pathogenesis of psoriasis will be introduced.

Many different components and processes contribute significantly to psoriasis disease initiation and maintenance (**Figure 6.1**). In the initiation phase, an unknown environmental factor induces keratinocytes to produce pro-inflammatory cytokines namely IL-1, IL-6 and Tumour Necrosis Factor (TNF). Under stress conditions, fibroblasts and innate immune cells such as the Natural Killer T (NKT) cells enhance local inflammatory response. Damaged keratinocytes release self DNA, which activates plasmocytoid dendritic cells (pDCs) leading to the release of IFN, which in turn activates dermal DCs. These dermal DCs migrate to the lymph nodes and present poorly characterised antigens to naive T-cells . Naive T cells usually travel around the body frequently visiting lymph nodes. Each of these naive T cells does so in the hope of finding a specific peptide which fits to its unique T cell Receptor (TCR). As a result of this interaction, naive T-cells proliferate and differentiate into CD4 Th1 (T helper 1) or Th17 (T helper 17) effector cells. Activated Th1 and Th17 express adhesion molecules and different chemokine receptors namely CXCR3, CCR4, and CCR6 on their surface. Due to the chemokines produced by activated keratinocytes, Th1 and Th17 migrate back to psoriatic dermis. In the inflammed skin, Th1 and Th17 secrete INF, IL-17, and IL-22. Effector cells are short-lived, but some activated CD4 and CD8 T-cells become memory cells. Cytokines produced by the Th1 and Th17 cells, stimulate keratinocytes to proliferate and produce phagocyte-recruiting chemokines such as CXCL1, CXCL3, CXCL5, CXCL8. In addition, activated DCs promote clonal expansion of memory CD4 and CD8 T-cells and their migration to the skin epidermis. This migration leads to better cross-talk between cells, and secretion of pro-inflammatory cytokines by CD4 (Th1, Th17) and CD8 (CTL) T-cells leads to amplification of the inflammatory reaction. As a result, specific psoriatic lesions occur (Nestle *et al.* [2009]).

6.2 Molecular Topology of Psoriasis Disease

Psoriatic human skin is a complex biological system (Nestle *et al.* [2009]) and a further investigation and characterisation of its genetic architecture might help reveal the biological basis of disease associations which modify disease progression, to ultimately provide insight into the pathogenesis of psoriasis. A systems-based

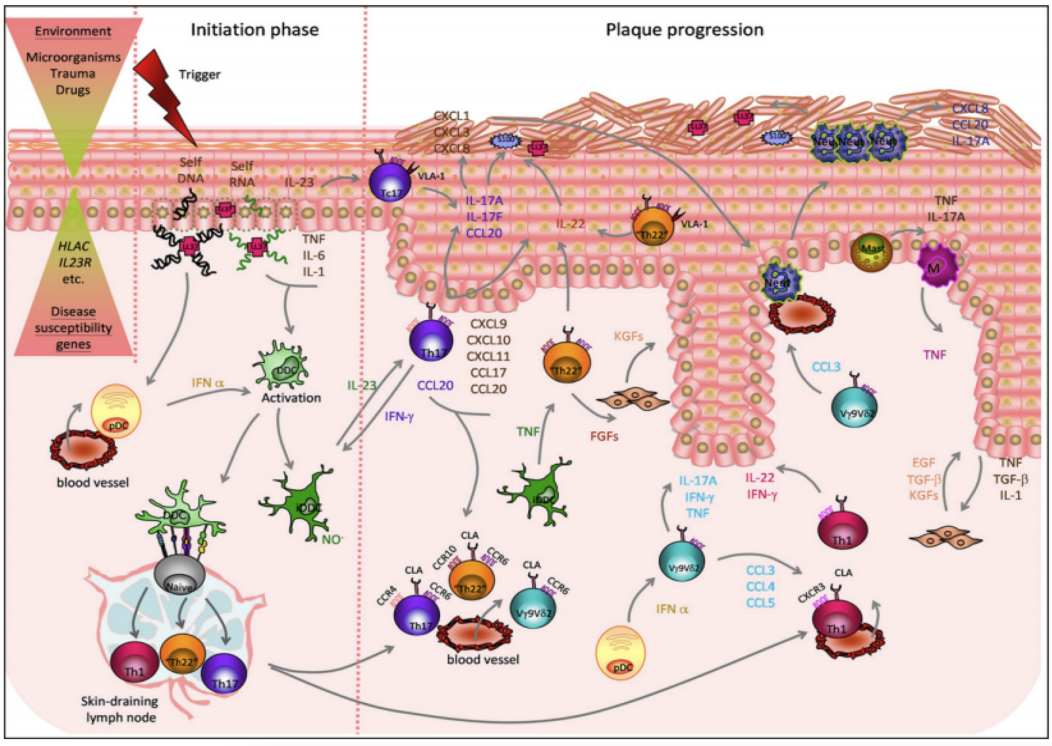

Figure 6.1: Immunopathogenesis of psoriasis (Picture adapted from Di Meglio *et al.* [2011])

analysis may be of value in attempting to provide a more comprehensive view of skin pathology. In recent years, data-driven approaches have been applied in understanding the pathogenesis and treatment of psoriasis (Gudjonsson *et al.* [2009b], Gudjonsson *et al.* [2010], Suarez-Farinas *et al.* [2010]). Some studies have attempted to elucidate the molecular pathways underlying in psoriasis, and other detailed gene expression studies have identified various differentially expressed genes by comparing non-lesional and lesional skin against normal tissue (Yao *et al.* [2008], Gudjonsson *et al.* [2010], Suarez-Farinas *et al.* [2010]). More recently, genome-wide expression analysis has been implemented to develop a computational pipeline to characterise heterogeneity among lesional skin samples (Ainali *et al.* [2012]), and also to identify inflammatory and cytokine activity phenotypes associated with presentation of chronic plaque psoriasis (Swindell *et al.* [2012]). Subsequently, initial computational models (Valeyev *et al.* [2010], Penner *et al.* [2012]) have focused on a dynamical systems approach, and have provided simplified but informative pictures of abhorrent cytokine interactions driving immune system responses characteristic of psoriasis. These modelling approaches are very useful for understanding general principles of how genetic perturbations may lead to psoriasis in a qualitative way.

However despite the recent considerable interest, a major challenge in translational research is the development of robust pipelines to modulate disease in an individualised and optimised fashion. Systems biology approaches that model biological systems can represent the complex disease interactions efficiently, and offer insights into underlying mechanisms and perturbations due to disease (Schadt & Bjorkegren [2012]). Since microarray-based gene expression data have been generated widely and contain information about concerted changes in transcript levels, an integrative network biology multistep procedure is proposed for identifying potential key drivers of psoriasis. Using gene expression profiles to infer gene-gene interactions, a gene co-expression network is generated. With graph-clustering techniques (Newman [2006]), we further identify clusters of genes and determine cytokine relevance. Because clusters of co-expressed genes are known to be functionally related (Stuart *et al.* [2003]), functional enrichment for GO or KEGG pathway terms is performed. In that way, this analysis not only identifies known biological processes and pathways involved in psoriasis disease, but also

offers a better understanding of the molecular basis of this disease phenotype by establishing cytokine-related key networks of co-expressed genes, as well as their topological and functional properties.

6.2.1 Materials and Methods

For this study, the dataset described, previously in **chapter 5**, is used. Briefly, we used tissue samples from 71 individuals, of which 34 were healthy controls (NN) and 37 were biopsies from patients affected by chronic plaque psoriasis. Tissues from both lesional (PP) and non-lesional (PN) areas were taken. Raw data were normalized using quantile normalization and expression estimates were computed using the Robust Multichip Average (RMA) method, on the Affymetrix HU133 Plus 2.0 platform (Irizarry *et al.* [2003]).

Microarray expression profiles for normal human epidermal Keratinocytes (KCs) treated with IL-20 family cytokines were also extracted from GEO database (*GSE7216*) (Sa *et al.* [2007]). Also, expression profiles for normal human epidermal Keratinocytes (KCs) treated with IL-17 were derived (*GSE7216*). Differential Expression Analysis was carried out using the MATLAB software and functions from Bioinformatics Toolbox [1]. Genes that were not expressed or had only small variability across the treated and untreated samples were removed. A t-test was conducted to identify significant changes for each gene in their expression values between the different phenotypes so as to extract cytokine-related over and/or under expressed genes (Dudoit *et al.* [2003]). Genes with $p - value < 0.05$ and $foldchange(fch) > 1.5$ were considered to be statistically significant.

Gene networks were inferred by using Pearson Correlation Coefficient (PCC), one of the metrics that were discussed in **Chapter 4**. Pairwise Pearson Correlation Coefficient (PCC) is estimated across the gene expression profiles and a similarity matrix was calculated for each of the skin tissue types (NN, PN and PP). Gene pairs that correlated above a threshold of 0.7 PCC were represented in the form of a network and a co-expression network was visualised using the Cytoscape software. Nodes (vertices) corresponded to genes and edges (links) corresponded to the similarity of the expression of two genes were across different

[1] http://www.mathworks.co.uk/products/bioinfo/

samples. Efficiency and stability of the topology of our networks was examined by rewiring links in the original networks (Maslov & Sneppen [2002]). For functional enrichment analysis of gene sets, the DAVID 6.7 bioinformatics database (June 2011) [1] was used and genes were matched with Gene Ontology (GO) terms and KEGG or Reactome pathways. Further to identify cluster structure in our networks, we used Markov Cluster Algorithm (MCL) (Enright *et al.* [2002]), which is also considered in **Chapter 5**.

6.2.2 Co-expression Models of Skin-related Phenotypes

Co-expression networks of skin phenotypes, normal, peri-lesional and lesional tissue samples, were built using all the genes at PCC 0.9, followed by a power law distribution and R-squared values ranging between 0.7 to 0.95, suggesting the structure of scale free networks (**Figure 6.2**). Highly connected nodes, hubs, indicate the likelihood that a gene is essential and its protein product play a major role in activation or inhibition of other proteins. Note here, that R-squared value for peri-lesional phenotype (PP) is larger (R^2=0.914) indicating a stronger linear association between the data variables (genes). The cut-off value of PCC $\geq$ 0.9 was selected according to recent studies where they have employed similar threshold criteria and have shown that gene co-expression landscape derived from PCC over 0.7 were proved to be more biologically relevant (Elo *et al.* [2007], Prieto *et al.* [2008]). In addition, below this cut-off the networks are very large suggesting the potential existence of many false positive edges (**Figure 6.3**).

Local topological parameters of the identified networks were further estimated and presented in Table 6.1. Networks derived from skin phenotypes were highly structured with high connectivity (average degree range: 6.3 to 49.7) and low clustering coefficient (range: 0.24 to 0.49). Interestingly, NN network has a large number of nodes with high connectivity (hubs) whereas networks derived from the patient tissue samples (PN and PP) have less number of hubs as is inferred by the average degree values.

However, due to the cytokine environment that sustains the pathogenic cascade of the psoriasis disease, sets of cytokine-responsive transcripts, identified

[1] http://david.abcc.ncifcrf.gov/

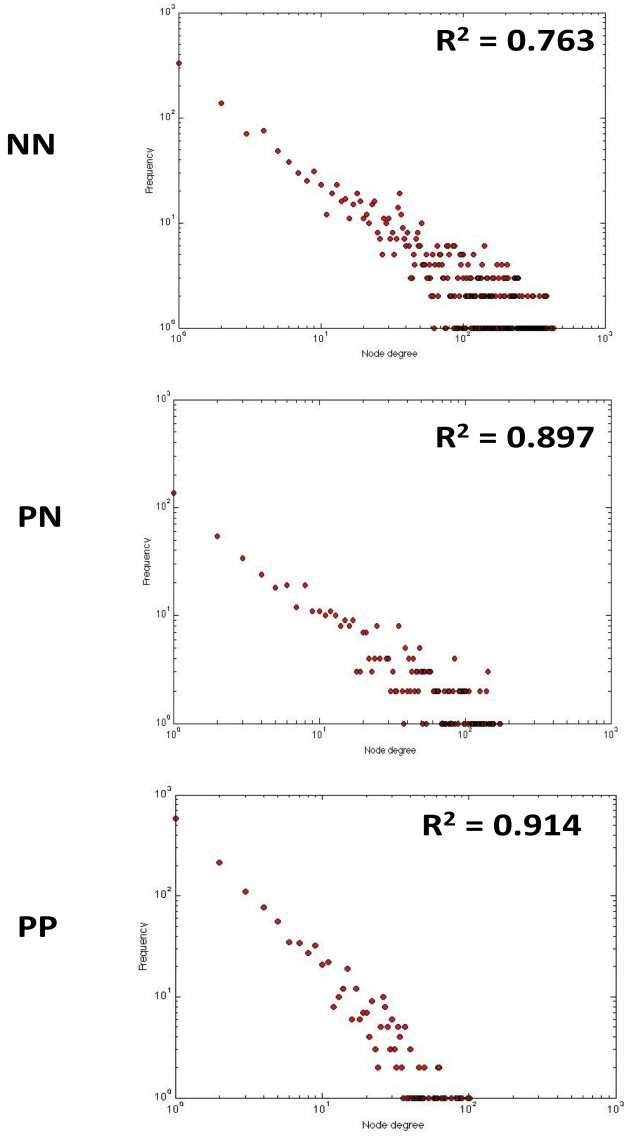

Figure 6.2: Log-log plots of node degree and frequency distribution for the three skin phenotypes. The power law distribution suggests that all the networks are scale-free

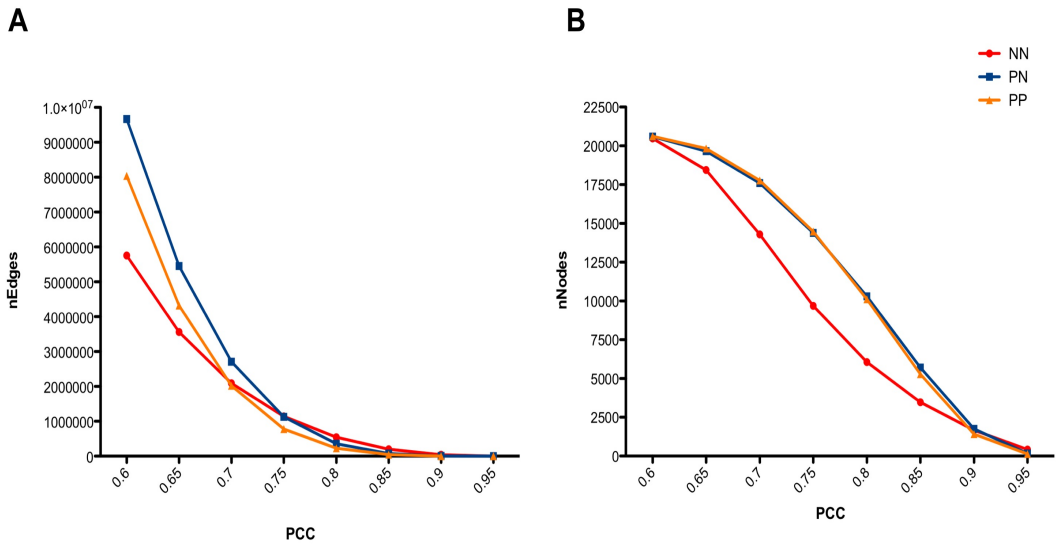

Figure 6.3: Association plots of (A) the number of edges/links included in the network according to the cut-off value of the pearson correlation coefficient (PCC) and (B) the number of genes .

Table 6.1: Topological parameters of co-expression networks derived from the three skin phenotypes (NN, PN and PP, respectively)

Networks	Number of Nodes	Number of Edges	Average Degree	Diameter	Clustering Coefficient
NN	1664	41359	49.710	9	0.484
PN	1753	7326	8.358	24	0.2763
PP	1416	4461	6.3008	18	0.243

from in vitro experiments performed with cultured keratinocytes, are used to map onto our networks and identify differences in cytokine activity among the three skin phenotypes (NN, PN and PP) (Gudjonsson *et al.* [2004]). Molecular signatures are established for INFg, IL17, IL19, IL20, IL22, IL24, and IL1b cytokines and their existence in the different skin phenotypes is demonstrated (Table 6.2). IL17 is overrepresented in the networks with a high coverage in PP. IL22 and IL24 exhibit also higher coverage than the other cytokines in the peri-lesional (PP) network. Also, there is an increase in the presence of the cytokines from the PN to PP networks, demonstrating the potential induction of keratinocytes producing pro-inflammatory cytokines. The smallest effect has IL19. Intriguingly, there is a high coverage of IL10 family cytokines in normal (NN) network compared to the other skin phenotypes, which could be possibly explained due to the variability of healthy individuals.

Table 6.2: Presence of cytokines molecular signatures in the skin networks

Networks	INFg	IL17	IL19	IL20	IL22	IL24	IL1b
NN	1.368%	10.29%	0.64%	1.37%	15.6%	5.99%	2.32%
PN	0.7%	7.94%	0.26%	1.27%	8.33%	2.70%	1.60%
PP	1.07%	10.58%	0.54%	1.35%	6.88%	3.34%	2.48%

6.2.3 Critical modules in lesional skin exhibit cytokine-related sub-networks

MCL clustering is further implemented for identifying groups of genes that are highly connected in the skin networks. The algorithm finds cluster structure in graphs by estimating the random walks through it and determines clusters via genes sharing higher-order connectivity in their local neighborhoods (Enright *et al.* [2002]). Using the default settings of the algorithm, 194 clusters were found in the NN network, of which five contained more than 30 genes. Similarly, 419 clusters were found in PN with only three consisting of more than 20 genes and 362 clusters in PP network, of which six contained more than 25 genes. Enrichment of each largest MCL cluster in the PP network was further applied for GO and pathway terms and identified epidermis development, immune response, keratinocyte differentiation as well TGF signalling, Pi3k/Akt Signaling and other immune related pathways. We also observed an overrepresentation of IL22 and IL24 cytokines (blue and red, respectively, in **Figure 6.4**) in one of the largest networks that was enriched in epidermis development, which was recently validated (He & Liang [2010]) (**Figure 6.4**). IL17, as well, is represented in one of the largest networks.

Therefore, cytokines activity was identified within the psoriatic network and the importance of IL22 and IL24 in keratinocyte and epidermis development was ascertained. The observations highlight the role of cytokine factors in the prevention and treatment of psoriasis and pinpoint the need for further validation as therapeutic targets. In the following section, IL22 is applied in different whole tissue models to uncover its underlying role in the pathogenesis of the disease and explore the clinical relevance of it in psoriasis. Combining disease relevant models and integrative network analysis, the in vivo effect of IL-22 is studied on normal human skin and its kinetics in human psoriasis pathogenesis, and in disease progression.

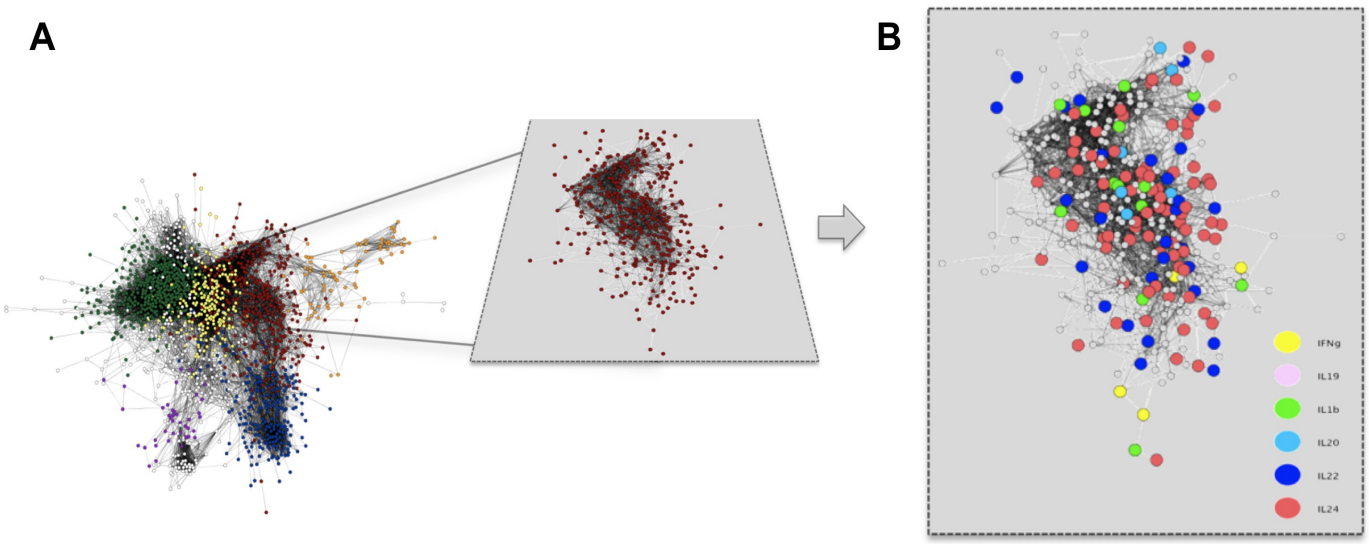

Figure 6.4: Clustering within psoriatic gene coexpression network to identify groups of genes that are functionally related (step 1). The largest clusters are visualized and cytokine - gene sets are aggregating within the largest cluster in the network (step 2).

6.3 Molecular profiling of in vivo models of human skin reveals IL-22 as a therapeutic target in the prevention and treatment of psoriasis

The discovery of new cytokines as therapeutic targets has great promise, but limitations abound due to the lack of clinical relevance when translating in vitro and animal data into the context of human disease (Chapman *et al.* [2007], Loisel *et al.* [2007]). Typically, preclinical models fail to reproduce the complexity of the human disease and a way of overcoming this shortfall is to develop *in vivo* models of disease using humanised mice. A xenotransplantation mouse model of human psoriasis has been previously established, where severely immunocompromised mice successfully engrafted human psoriatic tissue (Boyman & H. P. Hefti [2004]). Clinically relevant features of human psoriasis are reflected in the model including the complexity of genomic profiles operating within the transplanted skin grafts. Integration of these genomic profiles with publically available data leads to highly complex data sets. Deconvolution of these integrated data sets through network biology principles enables the formulation of new scientific hypotheses (Sirota & Butte [2011]) and the identification of unexpected mechanisms of disease pathogenesis as well potential therapeutic targets.

IL-22 is a key cytokine at sites of peripheral tissues, providing a link between the immune system and the epithelium (Wolk *et al.* [2004], Zheng *et al.* [2007]). Current in vitro and animal studies suggest that, in psoriasis, IL-22 is primarily produced by Th22 (Eyerich *et al.* [2009]) and Tc22 (intra-epidermal CD8+ cytotoxic T cells)(Res, Piskin et al. 2010) cells with a less significant contribution from Th17 and Th1 subsets (Boyman & H. P. Hefti [2004],Conrad *et al.* [2007]). Animal data suggest that IL-22 neutralisation is beneficial in inflammatory skin disease (Ma *et al.* [2008]). Also, the network biology cytokine-integrative approach described in **section 6.2**, has shown that IL-22 is a key player in psoriasis disease network. However, the precise function of IL-22 in human skin, its kinetics in disease pathogenesis and usefulness as therapeutic target remain unclear.

Human skin xenograft models were used to show that IL-22 is critical in the

development and progression of human psoriasis and a promising therapeutic target. A comprehensive framework for the discovery of biomarkers in human autoimmune disease was established (**Figure 6.5**), encompassing computational and experimental assessment of molecular interactions in whole tissue models of psoriasis. First, integrates clinical and genomic data, gathered from patients and tissue samples, to identify a candidate cytokine with a putative role in disease pathogenesis and therapy for further assessment (**Step 1**). Gene expression data pooled from large patient samples generate a disease interactome (network of inter-connected genes based on similar expression profiles), which is subsequently used to map gene expression data from experimental **Steps 2-5**. In **Step 2** the candidate cytokine is injected into a whole tissue model of normal human skin grafted onto immunosuppressed mice and the resultant pathology analyzed. **Step 3** generates a molecular signature of the cytokine-induced pathology in whole skin tissue. The cytokine molecular signature is mapped onto the disease interactome placing it in the context of human disease. **Step 4** reveals the kinetics of the cytokine in a whole tissue model of developing human disease and generates a molecular signature involved in disease development. At **step 5** the efficacy of anti-cytokine therapy is assessed in a clinically relevant model of severe disease and the anti-cytokine treatment molecular signature identified. At each step, molecular signatures provide a framework of potential biomarkers for testing in human disease as part of future clinical trials (dotted lines).

6.3.1 Materials and Methods

6.3.1.1 Data Sources

Normal human skin, from discarded plastic surgery specimens, and psoriasis samples, from patients attending dermatology clinics, were obtained in accordance with the Helsinki Declaration and approved by the Institutional review Board of Guys and St. Thomas NHS Foundation Trust Hospital. Written informed consent was obtained from all patients and healthy skin donors. Mice. The AGR129 mice, deficient in type I (A) and II(G) IFN receptors in addition to being Rag 2-/-, were kept pathogen free throughout the study and individually housed after experimentation. All animals were maintained and handled under an

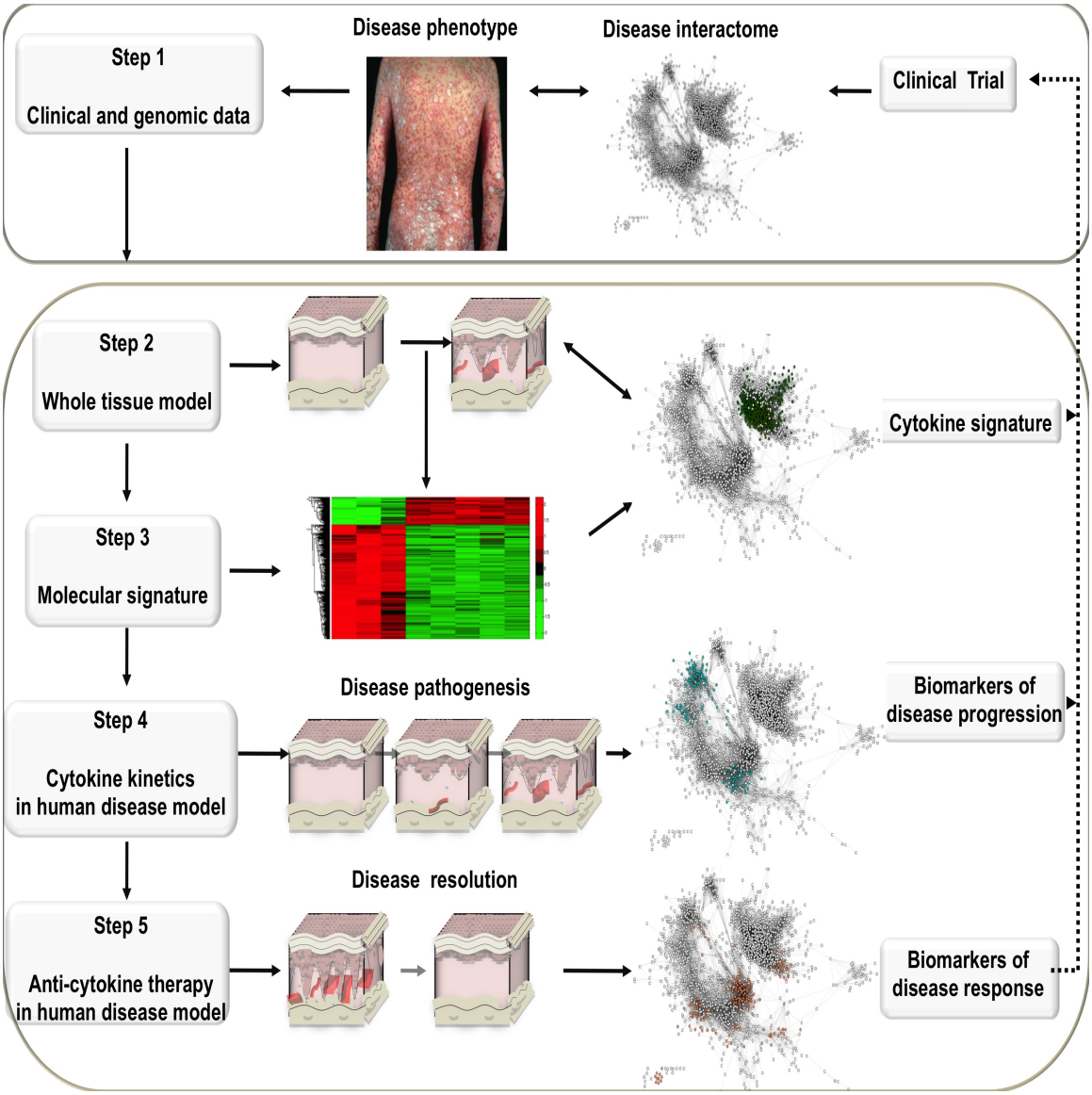

Figure 6.5: Biomarker Discovery Framework (BDF) in human autoimmune disease

approved license in accordance with the UK Home Office regulations for Animal Care. Transplantation of human skin (normal (NN) human skin and symptomless peri-lesional psoriasis(PN)) was performed onto AGR129 mice. Normal human (NN) skin was allowed to heal for two weeks prior to manipulation while PN skin was analyzed 14d and 35d after transplantation. The engrafted human skin was harvested at baseline (day 0), day 14 and day 35, after transplantation for time course assessment. A model of lesional psoriasis was developed by modifying the above model. A keratome biopsy from symptomatic lesional psoriasis (PP) skin was cut into 5mm x 5mm square pieces and glued onto the shaved dorsal aspects of the AGR129 mice. The skin was allowed to undergo wound healing for 1 week after which the tissue was manipulated. For analysis of the effect of neutralising studies the skin grafts were harvested 21 days after transplantation of lesional psoriasis skin.

Anti-IL-22 monoclonal antibodies were provided by Genentech and infliximab purchased from Centocor Ortho Biotec Inc. 1 week after transplantation of human lesional psoriasis skin, the anti-IL-22 moAbs were administered three times a week for the duration of 2 weeks and given at a dose of 150 g per mouse, whilst Infliximab was given, intraperitoneally (ip), at a dose of 100 g per mouse (5mg/kg), once every two weeks. Three independent experiments were performed with three individual human psoriasis skin donors. Three mice were used per time point.

The experimental design includes the following datasets: **1.** rhIL-22 was injected into normal human skin grafted onto AGR129 mice (NN-22), compared to skin grafts injected with control normal saline (NN-C). **2.** Human symptomless psoriasis (PN) skin was xenotransplanted onto AGR129 mice and analysed at different time points (from baseline baseline (day 0 - PNd0), day 14 (PNd14) and to psoriasiform skin by d35 (PNd35)). **3.** A model of lesional psoriasis was developed. PP skin at baseline (PPd0) had classical features of psoriasis including acanthosis and twenty-one days after transplantation of lesional psoriasis skin (PPd21), a more severe psoriatic phenotype developed. **4.** 1 week after transplantation of human lesional psoriasis skin (PP-C), anti-IL-22 monoclonal antibodies were administered three times a week for the duration of 2 weeks resulting in the PP-a22 phenotype, which was similar to PN phenotype from histological view (data not shown)

The number of samples that used for the analysis after the use of ANOVA testing to assess the statistical significant samples were: three controls (physiological saline, NN-C), five rhIL22 (recombinant IL-22), ten non-lesional psoriasis samples (PN) and ten lesional psoriasis samples (PP).

Experiments were done by Dr. Gayathri Perera at Guy's Hospital.

RNA integrity was assessed prior to performing expression arrays. Human Ref-6 Expression and Ref-8 BeadChips (Illumina, Ambion) were used to generate RNA expression data and were scanned with a BeadArray Reader (Illumina, San Diego, CA).

6.3.1.2 Methods

Differential gene expression analysis was carried out using the MATLAB software and functions from Bioinformatics Toolbox (http://www.mathworks.com/help/toolbox/bioinfo/). Genes that were not expressed or had only small variability across the samples were removed. Students t-test was conducted to identify significant changes for each gene in their expression values between the different phenotypes. Genes with p-values < 0.05 and fold change (fch) > 1.5 were considered to be statistically significant. Over and under expressed genes were further analysed for enriched Gene Ontology (GO) terms and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways using the log-odds ratio. High ratios indicate higher relative abundance of a GO term or KEGG pathway compared to the reference set.

The GAIN dataset described in section 6.2.1 was used to build a psoriasis network. The network was generated based on the co-expression of genes differentially expressed (p-value < 0.05 and fold change > 1.5) in the skin of patients with psoriasis (PP), compared to those in normal (NN) skin.Gene pairs that correlated above the 0.7 PCC threshold value were represented in the form of a network. This psoriasis co-expression network was used as a disease-relevant reference data set, against which the differentially expressed gene transcripts generated from the rhIL-22 injected model were assessed.

For functional enrichment analysis of gene sets, the DAVID 6.7 (http://david.abcc.ncifcrf.gov/) bioinformatics database (June 2011) was used and genes were matched with GO terms and KEGG pathway terms. The differentially expressed gene list was

submitted for functional classification in order to explore functionally related genes together. Functional annotation chart was used to provide gene-term over-representation analysis and to map the most relevant biological terms associated with a given gene list. Enriched biological annotation was chosen according to a statistical threshold.

6.3.2 A tissue specific IL-22 molecular signature

IL-22 is a key cytokine linking cells of the immune system with epithelial homeostasis and barrier function (Ouyang *et al.* [2011]). To elucidate the in vivo effects of IL-22 in human skin, rhIL-22 was injected into normal human skin grafted onto AGR129 mice (NN-22), compared to skin grafts injected with control normal saline (NN-C). NN-22 showed a remarkable phenotype dominated by epidermal thickening (acanthosis), replicating the epithelial changes seen in the inflammatory skin pathology psoriasis (**Figure 6.6a**). Increased acanthosis (**Figure 6.6b**) was accompanied by changes in psoriasis-associated markers and expression of molecular markers associated with psoriasis, such as S100A7 (psoriasin), serpin B4 and LCN2 (**Figure 6.6c**). Quantitative real time (RT) rt-PCR showed no significant change in TNF mRNA expression. IL-17A, IFN-g, IL-19 and IL-20 mRNA were not detected. RNA microarray analysis showed 4251 differentially expressed genes between NN-22 and NN-C. Hierarchical clustering of these genes identified a distinct set of genes associated with IL-22 injected human skin in vivo (**Figure 6.6d**), which we designated as the IL-22 *in vivo* molecular signature.

6.3.3 IL-22 has novel in vivo properties

A robust molecular network of human psoriasis was generated using patient data provided through the Genetic Association Information Network (GAIN) (Nair *et al.* [2009]), as described in the methods section. Here, the altered mechanisms of the disease comparing with the normal skin tissues were examined. The co-expression network was generated by assessing pairwise similarity of gene expression vectors as expressed by the Pearson correlation coefficient (PCC). Gene pairs that correlated above the 0.7 PCC threshold value were represented in the

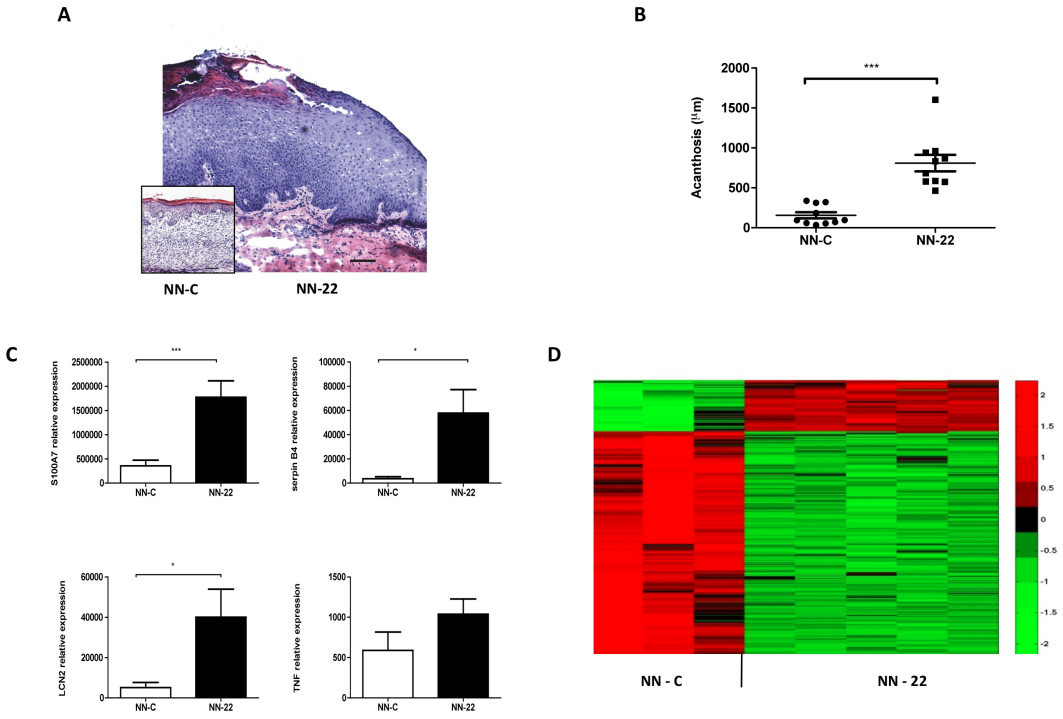

Figure 6.6: A tissue specific IL-22 molecular signature. (A) Histological analysis of NN skin grafted onto AGR129 mice after treatment with rhIL-22 (NN-22), or (inset) control (physiological saline, NN-C). H E staining shows increased epidermal thickness after rhIL-22 injection (scale bars 100m). Data representative of 10 mice with grafts from 3 independent normal human skin donors. (A) Assessment of epidermal thickness (acanthosis), in m, of NN-22 and NN-C xenografts. Each data point represents the mean of 10 random fields from 3 cryosections per mouse, *** $p<0.0001$. (C) Quantitative real time PCR analysis of S100A7, Serpin B4, LCN2 and TNF in NN-22 and NN-C xenografts, *$p<0.05$, *** $p<0.0001$. (D) Heatmap showing hierarchical clustering of differentially expressed genes between NN-22 and NN-C xenografts identifying a distinct in vivo IL-22 molecular signature. Red represents over-expressed genes and green represents under-expressed genes (Perera, GK., Ainali, C., *et al.* In Press)

form of a network. The resulting molecular network displayed 1614 differentially expressed genes (nodes) with 25281 interactions (edges). Nodes correspond to genes and edges represent similar expression profiles of linked genes across different samples (Vidal *et al.* [2011]). The psoriasis co-expression network (termed the GAIN-psoriasis molecular network) was used as a disease-relevant molecular reference data set, against which the differentially expressed gene transcripts generated by our cytokine injected whole tissue models were evaluated.

The NN-22 molecular signature was assessed within the GAIN-psoriasis molecular network in order to estimate its relevance to the disease. Of the 4251 genes differentially expressed between NN-C and NN-22, 421 genes mapped onto the GAIN-psoriasis molecular network (coverage of 26%) (**Figure 6.7a**). Functional enrichment analysis of these 421 genes, using the DAVID 6.7 bioinformatics database (June 2011) and matching the genes with KEGG and Gene Ontology (GO) terms (Huang *et al.* [2009]), identified genes involved in epidermal remodeling, immune response and Jak/Stat signaling, in keeping with the known functions of IL-22 (Ouyang *et al.* [2011]) (**Figure 6.7b**). Unexpectedly, there was a set of genes enriched for vascular remodeling (**Figure 6.7b**), which was confirmed using human CD31 staining of endothelial cells (data not shown).

6.3.4 Kinetics of IL-22 in psoriasis pathogenesis

Human symptomless psoriasis (PN) skin was xenotransplanted onto AGR129 mice and analysed at different time points. Histology (**Figure 6.8a**) confirmed the evolution of PN skin from baseline (PNd0) to psoriasiform skin by d35 (PNd35), as previously published (Boyman & H. P. Hefti [2004], Conrad *et al.* [2007]). There was also a significant increase in the number of CD31 staining cells from PNd0 to PNd14 and PNd35. Unsupervised Principle Component Analysis (PCA) showed clustering of differentially expressed genes in PN grafts at d14 (PNd14) and grafts at PNd35 compared to baseline PNd0 skin (**Figure 6.8b**). There was a clear separation between symptomless skin from evolving psoriasis as well as an emerging distinction between PNd14 (green dots) and PNd35 (blue dots), suggesting a similarity in the expression profiles in xenografts at day 14 and day 35. Next, differentially expressed genes from the *in vivo* IL-22 molecular sig-

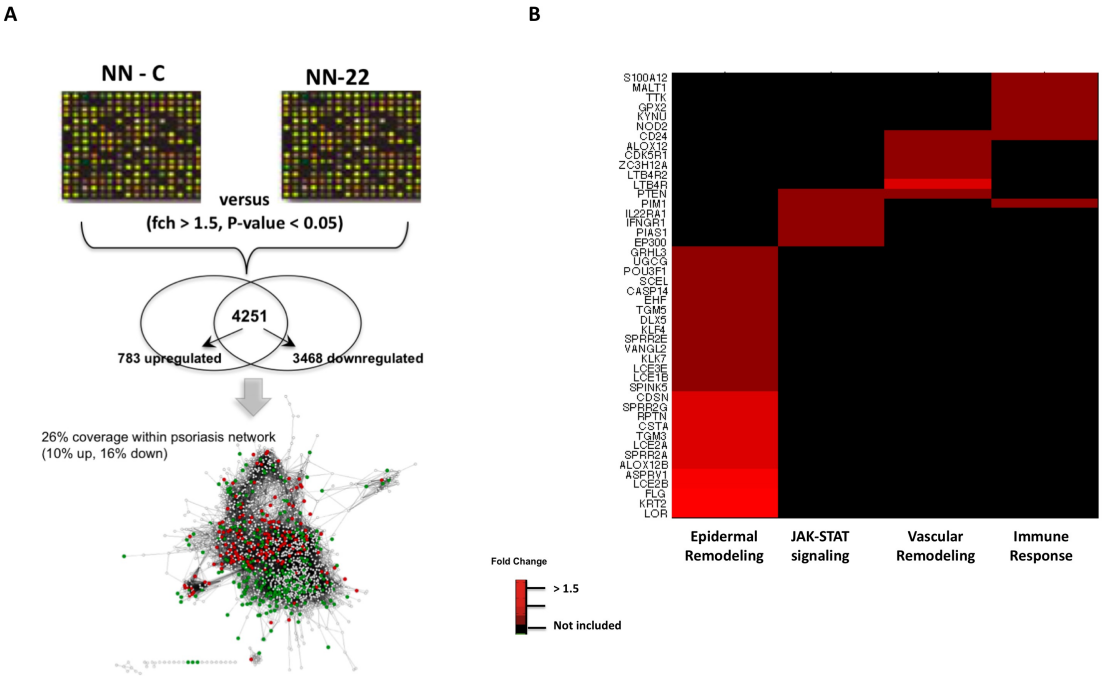

Figure 6.7: Mechanistic properties of IL-22 molecular signature. (A) The in vivo IL-22 molecular signature is shown using network analysis of differentially expressed genes between NN-C and NN-22 grafts (4250 genes) overlaid onto the GAIN-psoriasis molecular network (26% coverage). Over expressed genes (red dots) and under expressed genes (green dots). (B) Functional gene enrichment analysis of the differentially expressed genes between saline injected (NN-C) and IL-22 injected (NN-22) xenografts. The overall enrichment contribution is specified as a percentage of the relative abundance of each term in a cluster (Perera, GK., Ainali, C., *et al.* In Press).

nature were examined within the evolving psoriatic molecular network, resulting in 28 and 42 genes at d14 and d35, respectively. Comparison of the the co-expressed genes from the peri-lesional psoriasis skin model and the in vivo IL-22 molecular signature with the GAIN-psoriasis co-expression network was further done. There was a temporal increase in the number of in vivo IL-22 molecular signature genes mapping onto evolving psoriasis genes at d14 (28 genes) and d35 (42 genes) (**Figure 6.8c** red (over-expressed genes) and green (under-expressed genes) dots). Quantitative RT rt-PCR determined the temporal expression of IL-22 with a significant increase at d35 (**Figure 6.8d**), a relative expression of at least one order of magnitude greater than other IL-10 family cytokines (IL-19 and IL-20) and psoriasis-associated cytokines (IL-17A, IFN-). Functional gene enrichment analysis and subsequent hierarchical clustering showed genes involved in epidermal remodeling, keratinisation and response to wounding in PNd35 (**Figure 6.8e**), with a number of genes well known to psoriasis such as S100A7, Defb4, SPRR2A, serpinb4, serpinb3 and LCE3E.

6.3.5 An extreme phenotype of psoriasis is associated with a dominant IL-22 molecular signature

In the previous section was shown a critical role for IL-22 in developing psoriasis lesions. Further examination of IL-22 in established plaque psoriasis and in its progression to severe disease showed a dominant molecular signature. An in vivo model of lesional psoriasis was first establioshed and Human lesional psoriasis (PP) skin was transplanted onto immunosuppressed AGR129 mice. PP skin at baseline (PPd0) had classical features of psoriasis including acanthosis (**Figure 6.9a (i)**). Twenty-one days after transplantation of lesional psoriasis skin (PPd21), a more severe psoriatic phenotype developed (**Figure 6.9a(ii)**), exhibiting a significant increase in acanthosis (p ¡ 0.0001, Figure 6.9b), in the number of T cells ($p < 0.001$, **Figure 6.9b**). Differentially expressed genes between PPd0 and PPd21 were analysed to define the molecular characteristics of the PPd21 severe psoriasis phenotype. A total of 4363 genes were differentially expressed and 1142 were found to be in common with the in vivo IL-22 molecular signature(**Figure 6.9c**). Functional gene enrichment identified path-

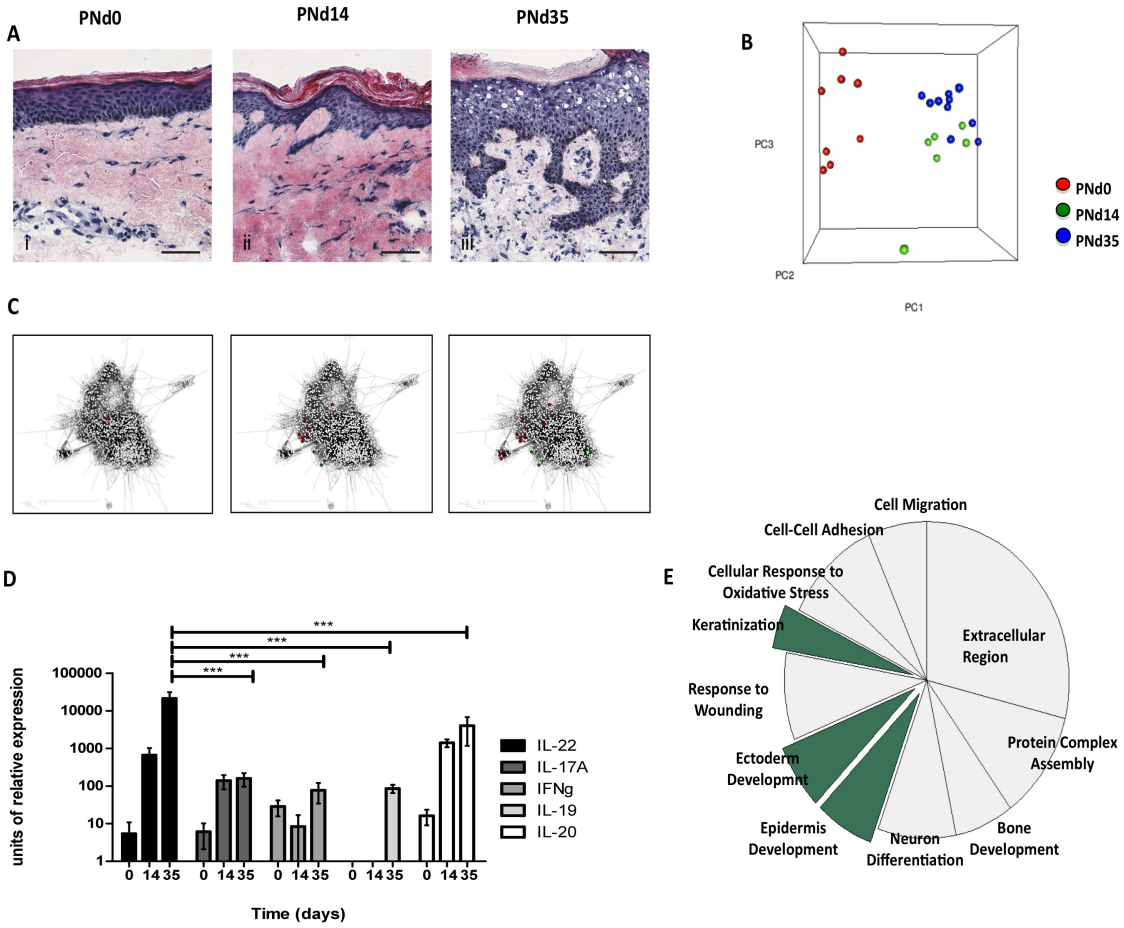

Figure 6.8: The kinetics of IL-22 in psoriasis pathogenesis. (A) Time course histological analysis of symptomless psoriasis skin (PN) grafts over 35 days. H E staining shows the temporal evolution of psoriasiform histology from baseline (PNd0) to day 14 (PNd14) and day 35 (PNd35) following skin transplantation (scale bars 100m). Data representative of 3 independent experiments using 3 different human donors. (B) Principle Component Analysis (PCA) showing the relationship between the common differentially expressed genes, and the clustering of samples at PNd0 (red dots), PNd14 (green dots) and PNd35 (blue dots). (C) Network analysis of differentially expressed genes in PNd0, PNd14 and PNd35, compared with the in vivo rhIL-22 molecular signature (red dots), overlaid onto the GAIN-psoriasis molecular network. (D) Time course quantitative real time rt-PCR analysis of cytokines in the evolution of psoriasis, (PN at days 0, 14 and 35), *$p<0.05$, **$p<0.001$, *** $p<0.0001$. Expressed as units of relative expression, normalized against GAPDH (E) Functional enrichment analysis of the genes playing a role in PNd35 (Perera, GK., Ainali, C., *et al.* In Press).

ways involved in epidermal remodeling, immune response pathways and Jak/Stat pathways (**Figure 6.9d**). Vascular remodeling, similar to that obtained with the in vivo IL-22 molecular signature, was also represented in PPd21 grafts. The IL-22 gene expression profile of PPd21 was distinct from the PNd35 gene expression profile with significant representation of IL-22 induced genes associated with vascular remodeling, immune response and Jak/Stat signaling.

6.3.6 Anti-resposive elemets of psoriasis treatment

Genes differentially expressed in anti-IL-22 treated psoriasis xenografts, IL-22 injected skin xenografts and the GAIN psoriasis molecular signature were integrated. 1164 genes were differentially expressed between PP-C and PP-a22 with 370 (327 plus 43) genes in common with the in vivo IL-22 signature and 43 of these mapped onto the GAIN-psoriasis molecular network (**Figure 6.10**). Functional enrichment of the 43 IL-22 response genes showed under expression of numerous genes involved in immune response and the Jak/Stat pathway, a critical component of the IL-22 signaling pathway. 19 genes, that were highly expressed in NN-22 and PPd21, were also down-regulated after anti-IL-22 treatment (**Figure 6.11a**, see black box). The PIM1 gene stood out with a low p value in the IL-22 signature (p=0.00249) and association with the IL-22 and psoriasis relevant Jak/Stat signaling pathway. Pim-1 had also been associated with vascular remodeling, mainly in the context of arterial hypertension (Meloche et al., 2011). Network topology of PIM1 revealed connection with STAT3 and S100A7 (**Figure 6.10b**), suggestive of a possible functional interaction of these 3 genes. To further determine the role of PIM1, in vitro assays using cytokine-treated primary skin cells were performed. PIM1 is known to be expressed by a variety of cell types including keratinocytes (Stewart and Rice, 1995), vascular smooth muscle cells (VSMCs)(Katakami et al., 2004). Based on experimental and network topology data, PIM1 is potentially induced indirectly in vascular cells via a mechanism involving S100A7 and its receptor RAGE (Receptor for Advanced Glycation Endproducts).

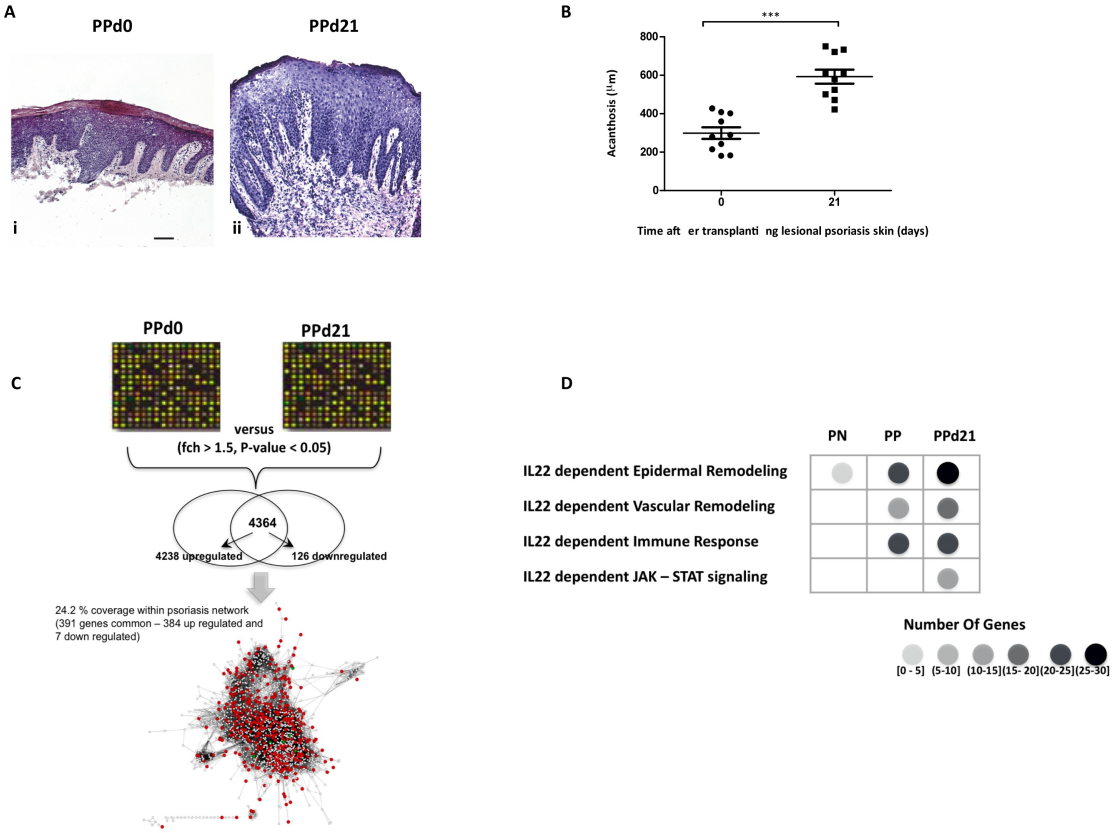

Figure 6.9: A severe psoriasis model. (A) HE staining of (i) PP at day 0 (PPd0), (ii) isotype injected PP at day 21 (PPd21) displaying a severe psoriasis phenotype. (B) Quantitative histological assessment of epidermal thickness, acanthosis, measured in m, in xenografts at PPd0 and PPd21. Each data point represents the mean of 10 measures from 3 random fields per mouse, ** $p<0.05$, *** $p<0.001$. (C) The IL-22 related severe psoriasis molecular signature mapped onto the GAIN-psoriasis molecular network, with over expressed (red dots) and under expressed genes (green dots). (D) Functional enrichment analysis of severe psoriasis IL-22 molecular signature (Perera, GK., Ainali, C., *et al.* In Press).

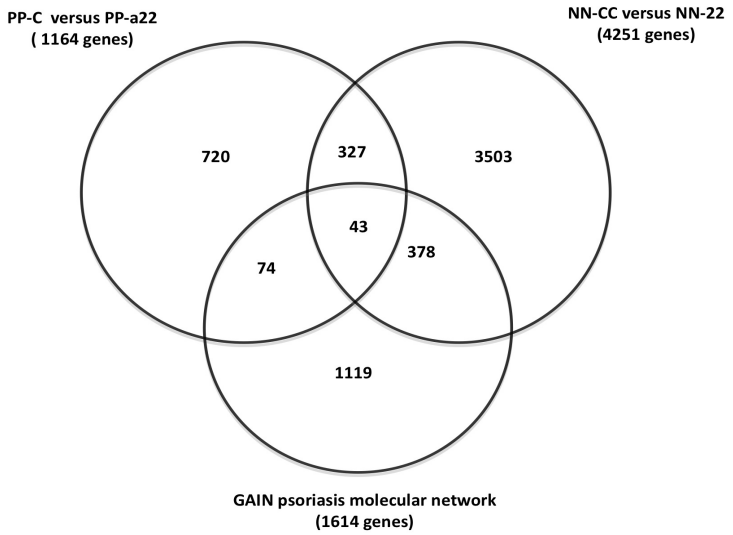

Figure 6.10: Venn Diagram of overlapping genes among anti-IL-22 (PP-a22), the in vivo IL-22 (NN-22) and the GAIN psoriasis molecular signature identifies 43 genes (Perera, GK., Ainali, C., *et al.* In Press).

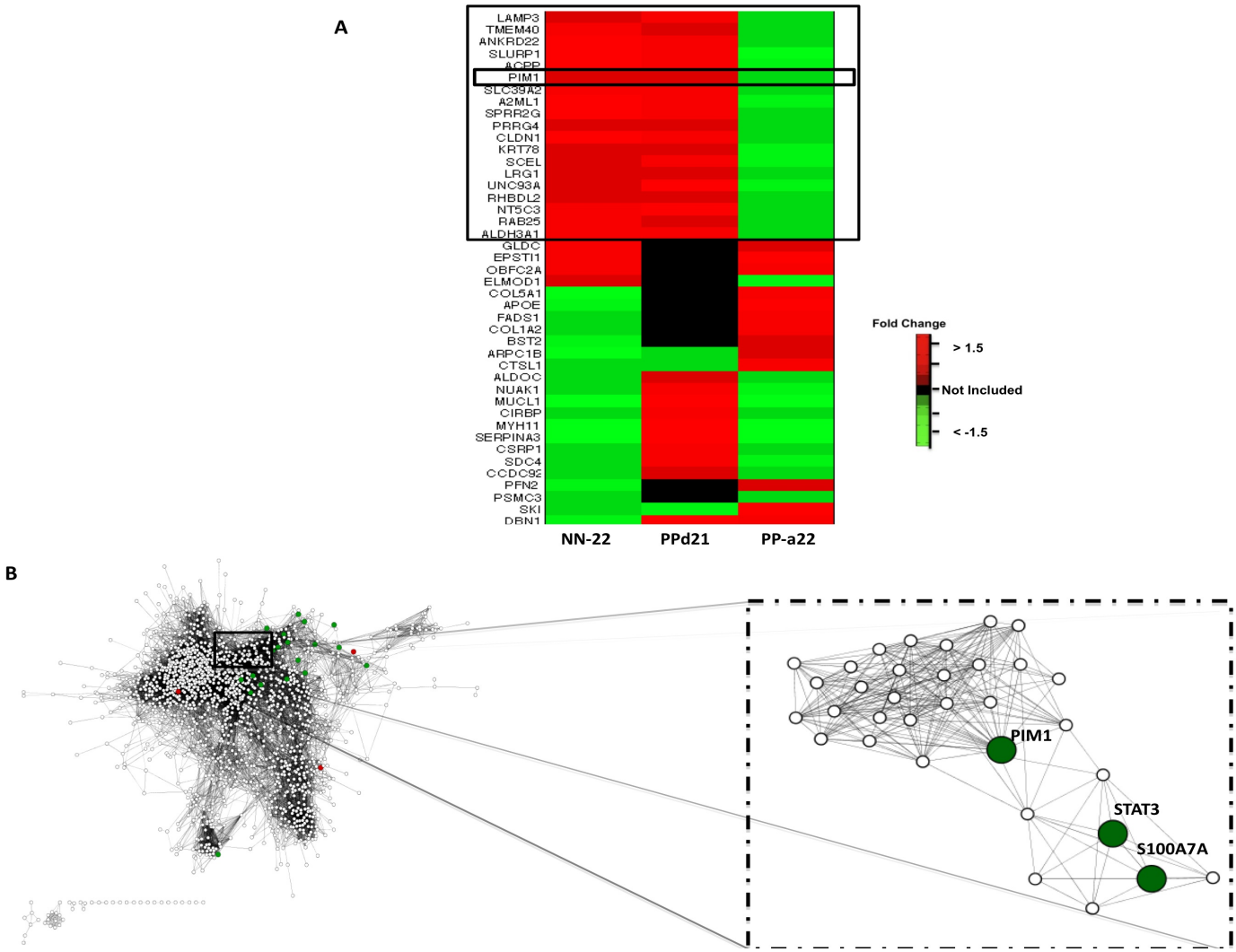

Figure 6.11: An IL-22 response gene signature identifies PIM1 as an important molecular checkpoint.(A) Heatmap showing gene expression of the IL22 response molecular signature, expressed as fold change. The black box identifies genes upregulated in NN-22 and PPd21 and down-regulated upon anti-IL-22 therapy (PP-a22) (B) Network topology analysis of PIM1 within the GAIN molecular network highlights the S100A7A STAT3 PIM1 connectivity (broken line box) (Perera, GK., Ainali, C., *et al.* In Press).

6.4 Discussion

Recent progress in characterizing human disease at a whole genome level offers the opportunity to utilise bioinformatics tools to explore disease pathogenesis (Conrad *et al.* [2007]). In **section 6.2**, molecular topology of psoriasis phenotype was examined and cytokine-related sub-networks were determined. The molecular gene-gene network was constructed using the most comprehensive gene expression data set of psoriasis to date publicly available from the GAIN study (Nair *et al.* [2009]). The GAIN-psoriasis molecular network served as a disease-relevant molecular reference data set for the experimental whole tissue model gene expression data. IL-20 family cytokines play a major role in psoriatic skin comparing to normal skin, with an emergence of IL22 and IL24. These findings support the role of cytokines, as successful disease targets and established relevance in autoimmunity (Ouyang *et al.* [2011].

IL-22 was chosen as a candidate cytokine for further experiments, due to its function as a key effector cytokine linking immune and epithelial cells and emerging data on its role in psoriasiform skin inflammation (Nawijn *et al.* [2011]). Human tissue models were further developed and a novel integrative framework for biomarker discovery was proposed. As a result, novel cytokine functions and therapeutic roles in psoriasis were discovered and serine/threonine kinase Pim-1 identified as a mechanistically important molecular checkpoint.

One of the unexpected outcomes of our network analysis of IL-22 injected human skin was the discovery of a prominent vascular remodeling gene expression signature. This finding was experimentally validated by demonstrating expanded and tortuous blood vessels in IL-22 injected human skin grafts. The morphology of IL-22 induced human blood vessels had striking similarities with those found in psoriatic skin (Braverman & Yen [1977a], Braverman & Yen [1977b]) and synovia of psoriatic arthritis (Reece *et al.* [1999]).

These results highlight the different roles of IL-22 at different stages of psoriasis development. An integrative network biology approach was applied to *in vivo* models of human disease, paving the way for generating "cytokinomes" of potential biological and therapeutic relevance for human autoimmune disease.

6.5 Disclaimer

The experiments were done by Dr. Gayathri Perera at Guy's Hospital. I am the sole contributor of the network analysis pipeline developed for data integration and biomarker discovery

Chapter 7

Conclusions and Future Research

The increasing flow of biomedical data and the advances in high-throughput experimental technologies present significant challenges in translational research. In translational medicine, scientists begin at "the bench" by studying disease at a molecular or cellular level and continue in the clinical level, at the "bedside". In bioinformatics, computational methods are developed to incorporate genotypic, phenotypic, and environmental knowledge to gain a better insight into complex diseases. Bringing together the computational genomics expertise and the biomedical research knowledge will provide important and novel insights into the mechanisms of disease pathophysiology.

This thesis has presented different approaches to make use of the biomedical data in order to effectively transform the information gained from newly generated data and to produce knowledge. The guiding theory for this research has been the development and application of machine learning strategies to facilitate the translational of findings from basic science and enhance human health.

At a first level, classification and feature selection techniques were studied to examine the prediction performance of a hybrid model and to identify promising candidate genes with respect to a phenotype of interest. Next, this thesis introduces a new integer optimisation classification algorithm for disease classification which for the first time was implemented on a "top-down" approach. The underlying algorithm was performed in a multi-stage procedure to reveal functional modules that are closely associated to the phenotype of interest and relevant to disease pathology. In that way, putative biomarkers, called Phenotype-

Responsive Genes (PRGs), were identified based on non-overlapping constraints of the classification procedure.

On another level, the methodological contributions of this thesis focused on the need for the development of diagnostic and prognostic tools as well personalized treatment of diseases. Psoriasis was used as an example disease and biologically meaningful phenotypic classes were generated and molecular variations among these groups were assessed. Using a tree-based approach, a new classification pipeline was developed to identify substantial phenotypic groups in psoriasis, based on patient gene expression profiles. The success of this framework has allowed the establishment of molecular patterns that characterize each sub-phenotype. This might uncover subtle differences in disease pathogenesis allowing the emergence of new treatments for psoriatic individuals and further facilitate the development of personalized treatments for the disease. However, further analysis and discovery of patterns and associations of transcripts of different cell-types (such as T-cells, dendritic cells, keratinocytes) must be done to shed light on the contribution that different cell types make towards the pathogenesis of psoriasis. We would then gain a better insight into this unique skin disease and hopefully, resolve some of the outstanding issues related to its pathogenesis and treatment.

However, the need to mine, visualise and integrate the different biological data has motivated the development of network biology (network medicine), which plays a major role in the understanding of the disease and the molecular mechanisms that drive it. Network inference approaches are valuable for the discovery of biomarkers associated with a disease. Here, using classification concepts, a network inference algorithm (NetCFS) was developed. The algorithm uses feature selection to select a number of genes that are highly correlated with the phenotype of interest so as to generate n different regression problems. The goal here is to identify sub-networks within the regulatory network that change between physiological and pathological stages of the disease. For this reason the algorithm was compared to co-expression network inferred be Pearson Correlation Coefficient and has been shown that can infer scale free networks exploiting static steady-state expression data and can identify hub genes that could be potential targets. Compared to GENIE3, this thesis applied the NetCFS to the

human genome and this raise some computational issues, due to the large dimensionality. Future advances of the algorithm should aim to improving significantly the network predictions and potentially applying it to time-series data as well genotyping data in order to capture the integrative perspective.

Due to an explosion in the search for putative disease biomarkers, during this work, some novel applications of network biology were employed for skin diseases; melanoma and psoriasis. An identification of biomarkers of human disease, and specifically those predictive of response to therapy, is an ideal of personalised medicine. In melanoma, exploiting protein-protein interaction data from public databases, the current study, described in **Chapter 4**, has shown that the newly generated IgE and IgG molecular signatures had different effect in tumor growth and were related to different pathways. Further, to explore the clinical relevance of IL-22 in psoriasis, a novel framework was implemented for interrogating cytokine pathways in human disease, by combining disease relevant models and integrative network analysis.Using a network biology approach, the in vivo effect of IL-22 has been shown on normal human skin and its kinetics in human psoriasis pathogenesis, and in disease progression. This integrative medicine systems analysis generated IL-22-related gene signatures, which could potentially identify subgroups of responders or not to anti-IL-22 therapy.

In this work, expression data were integrated with other biological public available data for integrative modeling. However, integrating genotypic data with gene expression may potentially associate genetic factors with the downstream changes in phenotype. For example, previously have been demonstrated that integration of Single Nucleotide Polymorphism (SNP) data enables the better understanding of the functional eect of SNPs in the structure and dynamics of biological networks (Bauer-Mehren *et al.* [2009]). Thus, trait-specific transcriptional networks of disease will provide the basis for comparative network analysis, where two networks relating to different disease states are handled simultaneously, in order to derive important changes unambiguously. Such analyses will reveal the critical gene variants in differentiating between disease groups or patient classes and might provide a predictive framework for cytokine (or cell-type) related pathway discovery.

References

Agrawal, e.a., M. A. Jade (2003). Support vector machines: a useful tool for process engineering applications. *Chem Eng Prog*, **99**, 57–62.

Ainali, C., Nestle, F., Papageorgiou, G.L. & Tsoka, S. (2011). Disease Classification through Integer Optimisation. *21st European Symposium on Computer Aided Process Engineering, AMSTERDAM:ELSEVIER SCIENCE BV*, 1548–1552.

Ainali, C., Valeyev, N., Perera, G., Williams, A., Gudjonsson, J.E., Ouzounis, C.A., Nestle, F.O. & Tsoka, S. (2012). Transcriptome classification reveals molecular subtypes in psoriasis. *BMC Genomics*, **13**, 472.

Alizadeh, A.A., Gentles, A.J., Alencar, A.J., Liu, C.L., Kohrt, H.E., Houot, R., Goldstein, M.J., Zhao, S., Natkunam, Y., Advani, R.H., Gascoyne, R.D., Briones, J., Tibshirani, R.J., Myklebust, J.H., Plevritis, S.K., Lossos, I.S. & Levy, R. (2011). Prediction of survival in diffuse large B-cell lymphoma based on the expression of 2 genes reflecting tumor and microenvironment. *Blood*, **118**, 1350–1358.

Ambesi-Impiombato, A., Bansal, M., Lio, P. & di Bernardo, D. (2006). Computational framework for the prediction of transcription factor binding sites by multiple data integration. *BMC Neurosci*, **7 Suppl 1**, S8.

Armutlu, P., Ozdemir, M.E., Uney-Yuksektepe, F., Kavakli, I.H. & Turkay, M. (2008). Classification of drug molecules considering their IC50 values using mixed-integer linear programming based hyper-boxes method. *BMC Bioinformatics*, **9**, 411.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Balch, C.M. & Soong, S.J. (2008). Predicting outcomes in metastatic melanoma. *J. Clin. Oncol.*, **26**, 168–169.

Bansal, M., Della Gatta, G. & di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, **22**, 815–822.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, **3**, 78.

Barabasi, A.L. (2005). Sociology. Network theory–the emergence of the creative enterprise. *Science*, **308**, 639–641.

Barabasi, A.L. & Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Bauer-Mehren, A., Furlong, L.I., Rautschka, M. & Sanz, F. (2009). From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways. *BMC Bioinformatics*, **10 Suppl 8**, S6.

Beer, M.A. & Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, **117**, 185–198.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. & Yakhini, Z. (2000). Tissue classification with gene expression profiles. *J. Comput. Biol.*, **7**, 559–583.

Ben-Hur, A., Ong, C.S., Sonnenburg, S., Scholkopf, B. & Ratsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Comput. Biol.*, **4**, e1000173.

Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J. & Sayers, E.W. (2012). GenBank. *Nucleic Acids Res.*, **40**, 48–53.

Botchkarev, V.A. (2003). Bone morphogenetic proteins and their antagonists in skin and hair follicle biology. *J. Invest. Dermatol.*, **120**, 36–47.

Bowcock, A.M., Shannon, W., Du, F., Duncan, J., Cao, K., Aftergut, K., Catier, J., Fernandez-Vina, M.A. & Menter, A. (2001). Insights into psoriasis and other inflammatory diseases from large-scale gene expression studies. *Hum. Mol. Genet.*, **10**, 1793–1805.

Boyman, O. & H. P. Hefti, e.a. (2004). Spontaneous development of psoriasis in a new animal model shows an essential role for resident T cells and tumor necrosis factor-alpha. *J Exp Med*, **199**, 731–736.

Braverman, I.M. & Yen, A. (1977a). Ultrastructure of the capillary loops in the dermal papillae of psoriasis. *J. Invest. Dermatol.*, **68**, 53–60.

Braverman, I.M. & Yen, A. (1977b). Ultrastructure of the human dermal microcirculation. II. The capillary loops of the dermal papillae. *J. Invest. Dermatol.*, **68**, 44–52.

Brazma, A. & Vilo, J. (2000). Gene expression data analysis. *FEBS Lett.*, **480**, 17–24.

Breiman, L., Friedman, J., Stone, C.J. & Olshen, R. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, 1st edn.

Butte, A.J. (2008). Translational bioinformatics: coming of age. *J Am Med Inform Assoc*, **15**, 709–714.

Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R. & Kohane, I.S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 12182–12186.

Camargo, A. & Azuaje, F. (2007). Linking gene expression and functional network data in human heart failure. *PLoS ONE*, **2**, e1347.

Cancer Genome Atlas, . (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Capon, F., Di Meglio, P., Szaub, J., Prescott, N.J., Dunster, C., Baumber, L., Timms, K., Gutin, A., Abkevic, V., Burden, A.D., Lanchbury, J., Barker, J.N., Trembath, R.C. & Nestle, F.O. (2007). Sequence variants in the genes for the interleukin-23 receptor (IL23R) and its ligand (IL12B) confer protection against psoriasis. *Hum. Genet.*, **122**, 201–206.

Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Pujar, A., Shearer, A.G., Travers, M., Weerasinghe, D., Zhang, P. & Karp, P.D. (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **40**, D742–753.

Cataisson, C., Pearson, A.J., Tsien, M.Z., Mascia, F., Gao, J.L., Pastore, S. & Yuspa, S.H. (2006). Cxcr2 ligands and g-csf mediate pkcalpha-induced intraepidermal inflammation. *J Clin Invest*, **116**, 2757–66.

Chandriani, S., Frengen, E., Cowling, V.H., Pendergrass, S.A., Perou, C.M., Whitfield, M.L. & Cole, M.D. (2009). A core myc gene expression signature is prominent in basal-like breast cancer but only partially overlaps the core serum response. *PloS one*, **4**.

Chapman, K., Pullen, N., Graham, M. & Ragan, I. (2007). Preclinical safety testing of monoclonal antibodies: the significance of species relevance. *Nat Rev Drug Discov*, **6**, 120–126.

Cheang, M.C., Voduc, D., Bajdik, C., Leung, S., McKinney, S., Chia, S.K., Perou, C.M. & Nielsen, T.O. (2008). Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin. Cancer Res.*, **14**, 1368–1376.

Chen, L., Xuan, J., Riggins, R.B., Clarke, R. & Wang, Y. (2011). Identifying cancer biomarkers by network-constrained support vector machines. *BMC Syst Biol*, **5**, 161.

Chen, X.W. & Liu, M. (2005). Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **21**, 4394–4400.

Chopra, P., Lee, J., Kang, J. & Lee, S. (2010). Improving cancer classification accuracy using gene pairs. *PLoS ONE*, **5**, e14305.

Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Chuang, L.Y., Yang, C.H., Wu, K.C. & Yang, C.H. (2011). A hybrid feature selection method for DNA microarray data. *Comput. Biol. Med.*, **41**, 228–237.

Citri, A. & Yarden, Y. (2006). EGF-ERBB signalling: towards the systems level. *Nat. Rev. Mol. Cell Biol.*, **7**, 505–516.

Conrad, C., Boyman, O., Tonel, G., Tun-Kyi, A., Laggner, U., de Fougerolles, A., Kotelianski, V., Gardner, H. & Nestle, F.O. (2007). Alpha1beta1 integrin is crucial for accumulation of epidermal T cells and the development of psoriasis. *Nat. Med.*, **13**, 836–842.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P. & Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–697.

Cun, Y. & Frohlich, H.F. (2012). Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics*, **13**, 69.

Dagliyan, O., Kavakli, I.H. & Turkay, M. (2009). Classification of cytochrome P450 inhibitors with respect to binding free energy and pIC50 using common molecular descriptors. *J Chem Inf Model*, **49**, 2403–2411.

Dagliyan, O., Uney-Yuksektepe, F., Kavakli, I.H. & Turkay, M. (2011). Optimization based tumor classification from microarray gene expression data. *PLoS ONE*, **6**, e14579.

Daruwala, R.S., Rudra, A., Ostrer, H., Lucito, R., Wigler, M. & Mishra, B. (2004). A versatile statistical analysis algorithm to detect genome copy number variation. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 16292–16297.

Davidson, E. & Levin, M. (2005). Gene regulatory networks. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 4935.

De Las Rivas, J. & Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.*, **6**, e1000807.

de Saizieu, A., Certa, U., Warrington, J., Gray, C., Keck, W. & Mous, J. (1998). Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. *Nat. Biotechnol.*, **16**, 45–48.

di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E. & Collins, J.J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.*, **23**, 377–383.

Di Meglio, P., Perera, G.K. & Nestle, F.O. (2011). The multitasking organ: recent insights into skin immune function. *Immunity*, **35**, 857–869.

Diaz-Uriarte, R. & Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Doyle, J.P., Dougherty, J.D., Heiman, M., Schmidt, E.F., Stevens, T.R., Ma, G., Bupp, S., Shrestha, P., Shah, R.D., Doughty, M.L., Gong, S., Greengard, P. & Heintz, N. (2008). Application of a translational profiling approach for the comparative analysis of CNS cell types. *Cell*, **135**, 749–762.

Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C. & Krawetz, S.A. (2003). Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Dudley, J.T. & Butte, A.J. (2009). Identification of discriminating biomarkers for human disease using integrative network biology. *Pac Symp Biocomput*, 27–38.

Dudoit, S. & Fridlyand, J.e.a. (2002). Comparison of discrimination methods for the classication of tumors using gene expression data. *Journal of the American Statistical Association.*, **97**, 77–87.

Dudoit, S., Gentleman, R.C. & Quackenbush, J. (2003). Open source software for the analysis of microarray data. *BioTechniques*, **Suppl**, 45–51.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J.M. (1999). Expression profiling using cDNA microarrays. *Nat. Genet.*, **21**, 10–14.

Dutkowski, J. & Gambin, A. (2007). On consensus biomarker selection. *BMC Bioinformatics*, **8 Suppl 5**, S5.

Dutkowski, J. & Ideker, T. (2011). Protein networks as logic functions in development and cancer. *PLoS Comput. Biol.*, **7**, e1002180.

Edgar, R., Domrachev, M. & Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178.

Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14863–14868.

Eisenberg, D., Marcotte, E., McLachlan, A.D. & Pellegrini, M. (2006). Bioinformatic challenges for the next decade(s). *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **361**, 525–527.

Elo, L.L., Jarvenpaa, H., Oresic, M., Lahesmaa, R. & Aittokallio, T. (2007). Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. *Bioinformatics*, **23**, 2096–2103.

Enright, A.J., Van Dongen, S. & Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Eyerich, S., Eyerich, K., Pennino, D., Carbone, T., Nasorri, F., Pallotta, S., Cianfarani, F., Odorisio, T., Traidl-Hoffmann, C., Behrendt, H., Durham, S.R., Schmidt-Weber, C.B. & Cavani, A. (2009). Th22 cells represent a distinct human T cell subset involved in epidermal immunity and remodeling. *J. Clin. Invest.*, **119**, 3573–3585.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J. & Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., Macgrogan, G., Bergh, J., Cameron, D., Goldstein, D., Duss, S., Nicoulaz, A.L., Brisken, C., Fiche, M., Delorenzi, M. & Iggo, R. (2005). Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene*, **24**, 4660–4671.

Feng, B.J., Sun, L.D., Soltani-Arabshahi, R., Bowcock, A.M., Nair, R.P., Stuart, P., Elder, J.T., Schrodi, S.J., Begovich, A.B., Abecasis, G.R., Zhang, X.J., Callis-Duffin, K.P., Krueger, G.G. & Goldgar, D.E. (2009). Multiple Loci within the major histocompatibility complex confer risk of psoriasis. *PLoS Genet.*, **5**, e1000606.

Ferlay, J., Parkin, D.M. & Steliarova-Foucher, E. (2010). Estimates of cancer incidence and mortality in Europe in 2008. *Eur. J. Cancer*, **46**, 765–781.

Fontana, J.M., Alexander, E. & Salvatore, M. (2012). Translational research in infectious disease: current paradigms and challenges ahead. *Transl Res*, **159**, 430–453.

Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I.H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.

Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Gabriele, L., Moretti, F., Pierotti, M.A., Marincola, F.M., Foa, R. & Belardelli, F.M. (2006). The use of microarray technologies in clinical oncology. *J Transl Med*, **4**, 8.

Gardner, T.S. & Faith, J.J. (2005). Reverse-engineering transcription control networks. *Phys Life Rev*, **2**, 65–88.

Gatza, M.L., Lucas, J.E., Barry, W.T., Kim, J.W., Wang, Q., Crawford, M.D., Datto, M.B., Kelley, M., Mathey-Prevot, B., Potti, A. & Nevins, J.R. (2010). A pathway-based classification of human breast cancer. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 6994–6999.

Gautreau, A., Barrat, A. & Barthelemy, M. (2009). Microdynamics in stationary complex networks. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 8847–8852.

Ge, H., Walhout, A.J. & Vidal, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.*, **19**, 551–560.

Geva, S. & Sitte, J. (1991). Adaptive nearest neighbor pattern classification. *IEEE Trans Neural Netw*, **2**, 318–322.

Glen, J.J. (2001). Classification Accuracy in Discriminant Analysis: A Mixed Integer Programming Approach. *The Journal of the Operational Research Society*, **52**, pp. 328–339.

Glen, J.J. (2003). An iterative mixed integer programming method for classification accuracy maximizing discriminant analysis. *Comput. Oper. Res.*, **30**, 181–198.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. & Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Graczyk, P.P. (2007). Gini coefficient: a new way to express selectivity of kinase inhibitors against a family of kinases. *J. Med. Chem.*, **50**, 5773–5779.

Gudjonsson, J.E., Johnston, A., Sigmundsdottir, H. & Valdimarsson, H. (2004). Immunopathogenic mechanisms in psoriasis. *Clin. Exp. Immunol.*, **135**.

Gudjonsson, J.E., Aphale, A., Grachtchouk, M., Ding, J., Nair, R.P., Wang, T., Voorhees, J.J., Dlugosz, A.A. & Elder, J.T. (2009a). Lack of evidence for activation of the hedgehog pathway in psoriasis. *J. Invest. Dermatol.*, **129**, 635–640.

Gudjonsson, J.E., Ding, J., Li, X., Nair, R.P., Tejasvi, T., Qin, Z.S., Ghosh, D., Aphale, A., Gumucio, D.L., Voorhees, J.J., Abecasis, G.R. & Elder, J.T. (2009b). Global gene expression analysis reveals evidence for decreased lipid biosynthesis and increased innate immunity in uninvolved psoriatic skin. *J. Invest. Dermatol.*, **129**, 2795–2804.

Gudjonsson, J.E., Ding, J., Johnston, A., Tejasvi, T., Guzman, A.M., Nair, R.P., Voorhees, J.J., Abecasis, G.R. & Elder, J.T. (2010). Assessment of the psoriatic transcriptome in a large sample: additional regulated genes and comparisons with in vitro models. *J. Invest. Dermatol.*, **130**, 1829–1840.

Gunther, E.C., Stone, D.J., Gerwien, R.W., Bento, P. & Heyes, M.P. (2003). Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 9608–9613.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, **46**, 389–422.

Haider, A.S., Duculan, J., Whynot, J.A. & Krueger, J.G. (2006). Increased JunB mRNA and protein expression in psoriasis vulgaris lesions. *J. Invest. Dermatol.*, **126**, 912–914.

Hall, M.A. & Smith, L.A. (1998). Practical feature subset selection for machine learning.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer, 2nd edn.

He, M. & Liang, P. (2010). IL-24 transgenic mice: in vivo evidence of overlapping functions for IL-20, IL-22, and IL-24 in the epidermis. *J. Immunol.*, **184**, 1793–1798.

Heidema, A.G., Boer, J.M., Nagelkerke, N., Mariman, E.C., van der A, D.L. & Feskens, E.J. (2006). The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *BMC Genet.*, **7**, 23.

Hua, S. & Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Huang, d.a.W., Sherman, B.T. & Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**, 44–57.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L. & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, **5**.

Hwang, T., Sicotte, H., Tian, Z., Wu, B., Kocher, J.P., Wigle, D.A., Kumar, V. & Kuang, R. (2008). Robust and efficient identification of biomarkers by classifying features on graphs. *Bioinformatics*, **24**, 2023–2029.

Ideker, T., Galitski, T. & Hood, L. (2001). A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, **2**, 343–372.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. & Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Jaakkola, T., Diekhans, M. & Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies. *Proc Int Conf Intell Syst Mol Biol*, 149–158.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. & Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.

Jensen, L.J. & Bateman, A. (2011). The rise and fall of supervised machine learning techniques. *Bioinformatics*, **27**, 3331–3332.

Jiang, R., Tang, W., Wu, X. & Fu, W. (2009). A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, **10 Suppl 1**, S65.

Jorda, M.A., Rayman, N., Valk, P., De Wee, E. & Delwel, R. (2003). Identification, characterization, and function of a novel oncogene: the peripheral cannabinoid receptor Cb2. *Ann. N. Y. Acad. Sci.*, **996**, 10–16.

Kaleta, C., Gohler, A., Schuster, S., Jahreis, K., Guthke, R. & Nikolajewa, S. (2010). Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis. *BMC Syst Biol*, **4**, 116.

Kanehisa, M. (2001). Prediction of higher order functional networks from genomic data. *Pharmacogenomics*, **2**, 373–385.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–280.

Kann, M.G. (2010). Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief. Bioinformatics*, **11**, 96–110.

Kapetanovic, I.M., Rosenfeld, S. & Izmirlian, G. (2004). Overview of commonly used bioinformatics methods and their applications. *Ann. N. Y. Acad. Sci.*, **1020**, 10–21.

Kaufman, L. & Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 99th edn.

Keseler, I.M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muniz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., Kaipa, P., Spaulding, A., Pacheco, J., Latendresse, M., Fulcher, C., Sarker, M., Shearer, A.G., Mackie, A., Paulsen, I., Gunsalus, R.P. & Karp, P.D. (2011). EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res.*, **39**, D583–590.

Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C. & Meltzer, P.S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Kim, S., Imoto, S. & Miyano, S. (2004). Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *BioSystems*, **75**, 57–65.

Kirschner, M.W. (2005). The meaning of systems biology. *Cell*, **121**, 503–504.

Kishibe, M., Bando, Y., Terayama, R., Namikawa, K., Takahashi, H., Hashimoto, Y., Ishida-Yamamoto, A., Jiang, Y.P., Mitrovic, B., Perez, D., Iizuka, H. & Yoshida, S. (2007). Kallikrein 8 is involved in skin desquamation in cooperation with other kallikreins. *J. Biol. Chem.*, **282**, 5834–5841.

Kodama, Y., Mashima, J., Kaminuma, E., Gojobori, T., Ogasawara, O., Takagi, T., Okubo, K. & Nakamura, Y. (2012). The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. *Nucleic Acids Res.*, **40**, 38–42.

Krueger, J.G. (2002). The immunologic basis for the treatment of psoriasis with new biologic agents. *J. Am. Acad. Dermatol.*, **46**, 1–23.

Lander, E.S., Linton, L.M., Birren, B. & et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armananzas, R., Santafe, G., Perez, A. & Robles, V. (2006). Machine learning in bioinformatics. *Brief. Bioinformatics*, **7**, 86–112.

Lebwohl, M. (2003). Psoriasis. *Lancet*, **361**, 1197–1204.

Lee, E., Trepicchio, W.L., Oestreicher, J.L., Pittman, D., Wang, F., Chamian, F., Dhodapkar, M. & Krueger, J.G. (2004). Increased expression of interleukin 23 p19 and p40 in lesional skin of patients with psoriasis vulgaris. *J. Exp. Med.*, **199**, 125–130.

Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T. & Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, **4**, e1000217.

Lee, W.P. & Tzou, W.S. (2009). Computational methods for discovering gene networks from expression data. *Brief. Bioinformatics*, **10**, 408–423.

Leung, K.S., Wong, K.C., Chan, T.M., Wong, M.H., Lee, K.H., Lau, C.K. & Tsui, S.K. (2010). Discovering protein-DNA binding sequence patterns using association rule mining. *Nucleic Acids Res.*, **38**, 6324–6337.

Levine, M. & Davidson, E.H. (2005). Gene regulatory networks for development. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 4936–4942.

Li, L., Weinberg, C.R., Darden, T.A. & Pedersen, L.G. (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131–1142.

Li, T., Zhang, C. & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, **20**, 2429–2437.

Liu, J.J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L. & Ling, X.B. (2005). Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics*, **21**, 2691–2697.

Liu, Y., Helms, C., Liao, W., Zaba, L.C., Duan, S., Gardner, J., Wise, C., Miner, A., Malloy, M.J., Pullinger, C.R., Kane, J.P., Saccone, S., Worthington, J., Bruce, I., Kwok, P.Y., Menter, A., Krueger, J., Barton, A., Saccone, N.L. & Bowcock, A.M. (2008). A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS Genet.*, **4**, e1000041.

Lockhart, D.J. & Winzeler, E.A. (2000). Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Loisel, S., Ohresser, M., Pallardy, M., Dayde, D., Berthou, C., Cartron, G. & Watier, H. (2007). Relevance, advantages and limitations of animal models used in the development of monoclonal antibodies for cancer treatment. *Crit. Rev. Oncol. Hematol.*, **62**, 34–42.

Lowes, M.A., Bowcock, A.M. & Krueger, J.G. (2007). Pathogenesis and therapy of psoriasis. *Nature*, **445**, 866–873.

Luo, W., Friedman, M.S., Shedden, K., Hankenson, K.D. & Woolf, P.J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, **10**, 161.

Lussier, Y.A. & Li, H. (2012). The rise of translational bioinformatics. *Genome Biol.*, **13**, 319.

Ma, H.L., Liang, S., Li, J., Napierata, L., Brown, T., Benoit, S., Senices, M., Gill, D., Dunussi-Joannopoulos, K., Collins, M., Nickerson-Nutter, C., Fouser, L.A. & Young, D.A. (2008). IL-22 is required for Th17 cell-mediated pathology in a mouse model of psoriasis-like skin inflammation. *J. Clin. Invest.*, **118**, 597–607.

Ma, S. & Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Brief. Bioinformatics*, **9**, 392–403.

Macchiarulo, A., Thornton, J.M. & Nobeli, I. (2009). Mapping human metabolic pathways in the small molecule chemical space. *J Chem Inf Model*, **49**, 2272–2289.

Madhamshettiwar, P.B., Maetschke, S.R., Davis, M.J., Reverter, A. & Ragan, M.A. (2012). Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Med*, **4**, 41.

Marbach, D., Prill, R.J., Schaffter, T., Mattiussi, C., Floreano, D. & Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 6286–6291.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. & Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7 Suppl 1**, S7.

Maslov, S. & Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, **296**, 910–913.

Matsuno, H., Doi, A., Nagasaki, M. & Miyano, S. (2000). Hybrid Petri net representation of gene regulatory network. *Pac Symp Biocomput*, 341–352.

Mazumder, A. & Wang, Y. (2006). Gene-expression signatures in oncology diagnostics. *Pharmacogenomics*, **7**, 1167–1173.

McDonough, C.W., Hicks, P.J., Lu, L., Langefeld, C.D., Freedman, B.I. & Bowden, D.W. (2009). The influence of carnosinase gene polymorphisms on diabetic nephropathy risk in African-Americans. *Hum. Genet.*, **126**, 265–275.

McKinney, B.A., Reif, D.M., Ritchie, M.D. & Moore, J.H. (2006). Machine learning for detecting gene-gene interactions: a review. *Appl. Bioinformatics*, **5**, 77–88.

Melvin, I., Ie, E., Kuang, R., Weston, J., Stafford, W.N. & Leslie, C. (2007). SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition. *BMC Bioinformatics*, **8 Suppl 4**, S2.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W. & Hamprecht, F.A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, **10**, 213.

Meyer, P.E., Kontos, K., Lafitte, F. & Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinform Syst Biol*, 79879.

Miller, A. (2002). *Subset Selection in Regression, Second Editon*. Chapman and Hall/CRC.

Miller, L.D., Smeds, J., George, J., Vega, V.B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E.T. & Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts

mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 13550–13555.

Nair, R.P., Duffin, K.C., Helms, C., Ding, J., Stuart, P.E., Goldgar, D., Gudjonsson, J.E. & et al. (2009). Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat. Genet.*, **41**, 199–204.

Nawijn, M.C., Alendar, A. & Berns, A. (2011). For better or for worse: the role of Pim oncogenes in tumorigenesis. *Nat. Rev. Cancer*, **11**, 23–34.

Nestle, F.O., Kaplan, D.H. & Barker, J. (2009). Psoriasis. *N. Engl. J. Med.*, **361**, 496–509.

Newman, M.E. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 8577–8582.

Niijima, S. & Okuno, Y. (2009). Laplacian linear discriminant analysis approach to unsupervised feature selection. *IEEE/ACM Trans Comput Biol Bioinform*, **6**, 605–614.

Nomura, I., Gao, B., Boguniewicz, M., Darst, M.A., Travers, J.B. & Leung, D.Y. (2003). Distinct patterns of gene expression in the skin lesions of atopic dermatitis and psoriasis: a gene microarray analysis. *J. Allergy Clin. Immunol.*, **112**, 1195–1202.

Notter, M., Willinger, T., Erben, U. & Thiel, E. (2001). Targeting of a B7-1 (CD80) immunoglobulin G fusion protein to acute myeloid leukemia blasts increases their costimulatory activity for autologous remission T cells. *Blood*, **97**, 3138–3145.

Oestreicher, J.L., Walters, I.B., Kikuchi, T., Gilleaudeau, P., Surette, J., Schwertschlag, U., Dorner, A.J., Krueger, J.G. & Trepicchio, W.L. (2001). Molecular classification of psoriasis disease-associated genes through pharmacogenomic expression profiling. *Pharmacogenomics J.*, **1**, 272–287.

Ouyang, W., Rutz, S., Crellin, N.K., Valdez, P.A. & Hymowitz, S.G. (2011). Regulation and functions of the IL-10 family of cytokines in inflammation and disease. *Annu. Rev. Immunol.*, **29**, 71–109.

Ouzounis, C.A. (2012). Rise and demise of bioinformatics? Promise and progress. *PLoS Comput. Biol.*, **8**, e1002487.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546–554.

Pang, H. & Zhao, H. (2008). Building pathway clusters from Random Forests classification using class votes. *BMC Bioinformatics*, **9**, 87.

Park, I., Lee, K.H. & Lee, D. (2010). Inference of combinatorial Boolean rules of synergistic gene sets from cancer microarray datasets. *Bioinformatics*, **26**, 1506–1512.

Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., Kurbatova, N., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rustici, G., Sharma, A., Williams, E., Adamusiak, T., Brandizi, M., Sklyar, N. & Brazma, A. (2011). ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–1004.

Patel, J.L. & Goyal, R.K. (2007). Applications of artificial neural networks in medical science. *Curr Clin Pharmacol*, **2**, 217–226.

Pavlidis, P., Furey, T.S., Liberto, M., Haussler, D. & Grundy, W.N. (2001). Promoter region-based classification of genes. *Pac Symp Biocomput*, 151–163.

Pawitan, Y., Bjohle, J., Amler, L., Borg, A.L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., Liu, E.T., Miller, L., Nordgren, H., Ploner, A., Sandelin, K., Shaw, P.M., Smeds, J., Skoog, L., Wedren, S. & Bergh, J. (2005). Gene expression profiling

spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.*, **7**, R953–964.

Pazos, F. & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.*, **14**, 609–614.

Penner, K., Ermentrout, B. & Swigon, D. (2012). Pattern formation in a model of acute inflammation. *SIAM Journal on Applied Dynamical Systems*, **11**, 629–660.

Perumal, S.S., Shanthi, P. & Sachdanandam, P. (2005). Therapeutic effect of tamoxifen and energy-modulating vitamins on carbohydrate-metabolizing enzymes in breast cancer. *Cancer Chemother. Pharmacol.*, **56**, 105–114.

Powell, J. (1998). Enhanced concatemer cloning-a modification to the SAGE (Serial Analysis of Gene Expression) technique. *Nucleic Acids Res.*, **26**, 3445–3446.

Prakash, A. & Tompa, M. (2005). Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.*, **23**, 1249–1256.

Prieto, C., Risueno, A., Fontanillo, C. & De las Rivas, J. (2008). Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PLoS ONE*, **3**, e3911.

Qi, Y., Bar-Joseph, Z. & Klein-Seetharaman, J. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, **63**, 490–500.

Rainer, Z., Eferl, R., Kenner, L., Florin, L., Hummerich, L., Mehic, D., Scheuch, H., Angel, P., Tschachler, E. & Wagner, E.F. (2005). Psoriasis-like skin disease and arthritis caused by inducible epidermal deletion of jun proteins. *Nature*, **437**, 369–375.

Ramirez-Montagut, T., Turk, M.J., Wolchok, J.D., Guevara-Patino, J.A. & Houghton, A.N. (2003). Immunity to melanoma: un-

raveling the relation of tumor immunity and autoimmunity. *Oncogene*, **22**, 3180–3187.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabasi, A.L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Reece, R.J., Canete, J.D., Parsons, W.J., Emery, P. & Veale, D.J. (1999). Distinct vascular patterns of early synovitis in psoriatic, reactive, and rheumatoid arthritis. *Arthritis Rheum.*, **42**, 1481–1484.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P. & Mesirov, J.P. (2006). GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.

Risau-Gusman, S. & Gordon, M.B. (2001). Statistical mechanics of learning with soft margin classifiers. *Phys Rev E Stat Nonlin Soft Matter Phys*, **64**, 031907.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D. & Brown, P.O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.

Ryoo, H.S. (2006). Pattern classification by concurrently determined piecewise linear and convex discriminant functions. *Comput. Ind. Eng.*, **51**, 79–89.

Sa, S.M., Valdez, P.A., Wu, J., Jung, K., Zhong, F., Hall, L., Kasman, I., Winer, J., Modrusan, Z., Danilenko, D.M. & Ouyang, W. (2007). The effects of IL-20 subfamily cytokines on reconstituted human epidermis suggest potential roles in cutaneous innate defense and pathogenic adaptive immunity in psoriasis. *J. Immunol.*, **178**, 2229–2240.

Saeys, Y., Inza, I. & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Sanchez Claros, C. & Tramontano, A. (2012). Detecting mutually exclusive interactions in protein-protein interaction maps. *PLoS ONE*, **7**, e38765.

Saveria, P., Mascia, F., Mariani, V. & Girolomoni, G. (2007). The epidermal growth factor receptor system in skin repair and inflammation. *Journal of Investigative Dermatology*, **128**, 1365–1374.

Sayadi, A., Briganti, L., Tramontano, A. & Via, A. (2011). Exploiting publicly available biological and biochemical information for the discovery of novel short linear motifs. *PLoS ONE*, **6**, e22270.

Schadt, E.E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature*, **461**, 218–223.

Schadt, E.E. & Bjorkegren, J.L. (2012). NEW: network-enabled wisdom in biology, medicine, and health care. *Sci Transl Med*, **4**, 115rv1.

Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E. & Davis, R.W. (1998). Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.*, **16**, 301–306.

Schneider, M.R., Antsiferova, M., Feldmeyer, L., Dahlhoff, M., Bugnon, P., Hasse, S., Paus, R., Wolf, E. & Werner, S. (2008). Betacellulin regulates hair follicle development and hair cycle induction and enhances angiogenesis in wounded skin. *J. Invest. Dermatol.*, **128**, 1256–1265.

Schramm, G., Surmann, E.M., Wiesberg, S., Oswald, M., Reinelt, G., Eils, R. & Konig, R. (2010). Analyzing the regulation of metabolic pathways in human breast cancer. *BMC Med Genomics*, **3**, 39.

Schuster, S., Fell, D.A. & Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, **18**, 326–332.

Schwarz, D.F., Konig, I.R. & Ziegler, A. (2010). On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics*, **26**, 1752–1758.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M. & Ideker, T. (2005). Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 1974–1979.

Shen, H.B. & Chou, K.C. (2006). Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, **22**, 1717–1722.

Shendure, J. (2008). The beginning of the end for microarrays? *Nat. Methods*, **5**, 585–587.

Shi, T., Seligson, D., Belldegrun, A.S., Palotie, A. & Horvath, S. (2005). Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Mod. Pathol.*, **18**, 547–557.

Shyamsundar, R., Kim, Y.H., Higgins, J.P., Montgomery, K., Jorden, M., Sethuraman, A., van de Rijn, M., Botstein, D., Brown, P.O. & Pollack, J.R. (2005). A DNA microarray survey of gene expression in normal human tissues. *Genome Biol.*, **6**, R22.

Simon, R., Mirlacher, M. & Sauter, G. (2005). Tissue microarrays. *Methods Mol. Med.*, **114**, 257–268.

Sirota, M. & Butte, A.J. (2011). The role of bioinformatics in studying rheumatic and autoimmune disorders. *Nat Rev Rheumatol*, **7**, 489–494.

Sohn, I., Owzar, K., George, S.L., Kim, S. & Jung, S.H. (2009). A permutation-based multiple testing method for time-course microarray experiments. *BMC Bioinformatics*, **10**, 336.

Soinov, L.A., Krestyaninova, M.A. & Brazma, A. (2003). Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biol.*, **4**, R6.

Somorjai, R.L., Dolenko, B. & Baumgartner, R. (2003). Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, **19**, 1484–1491.

Staiger, C., Cadot, S., Kooter, R., Dittrich, M., Muller, T., Klau, G.W. & Wessels, L.F. (2012). A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. *PLoS ONE*, **7**, e34796.

Statnikov, A., Wang, L. & Aliferis, C.F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, **9**, 319.

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. & Gilles, E.D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190–193.

Sterk, P., Kulikova, T., Kersey, P. & Apweiler, R. (2007). The EMBL Nucleotide Sequence and Genome Reviews Databases. *Methods Mol. Biol.*, **406**, 1–21.

Steuer, R., Kurths, J., Daub, C.O., Weise, J. & Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18 Suppl 2**, S231–240.

Strange, A., Capon, F., Spencer, C.C., Knight, J., Weale, M.E., Allen, M.H., Barton, A., Band, G., Bellenguez, C., Bergboer, J.G., Blackwell, J.M., Bramon, E., Bumpstead, S.J. & Casas, e.a., J. P. (2010). A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat. Genet.*, **42**, 985–990.

Strobl, C., Boulesteix, A.L., Zeileis, A. & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307.

Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Suarez-Farinas, M., Lowes, M.A., Zaba, L.C. & Krueger, J.G. (2010). Evaluation of the psoriasis transcriptome across different studies by gene set enrichment analysis (GSEA). *PLoS ONE*, **5**, e10247.

Subelj, L. & Bajec, M. (2011). Unfolding communities in large complex networks: combining defensive and offensive label propagation for core extraction. *Phys Rev E Stat Nonlin Soft Matter Phys*, **83**, 036103.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. & Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.

Sun, L.D., Cheng, H., Wang, Z.X., Zhang, A.P., Wang, P.G., Xu, J.H., Zhu, Q.X. & Zhou, e.a., H. S. (2010). Association analyses identify six new psoriasis susceptibility loci in the Chinese population. *Nat. Genet.*, **42**, 1005–1009.

Swindell, W.R., Xing, X., Stuart, P.E., Chen, C.S., Aphale, A., Nair, R.P., Voorhees, J.J., Elder, J.T., Johnston, A. & Gudjonsson, J.E. (2012). Heterogeneity of inflammatory and cytokine networks in chronic plaque psoriasis. *PLoS ONE*, **7**, e34594.

Symmans, W.F., Liu, J., Knowles, D.M. & Inghirami, G. (1995). Breast cancer heterogeneity: evaluation of clonality in primary and metastatic lesions. *Hum. Pathol.*, **26**, 210–216.

Tan, A.C. & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification. *Appl. Bioinformatics*, **2**, 75–83.

Thomas, R.S., Rank, D.R., Penn, S.G., Zastrow, G.M., Hayes, K.R., Pande, K., Glover, E., Silander, T., Craven, M.W., Reddy, J.K., Jovanovich, S.B. & Bradfield, C.A. (2001). Identification of toxicologically predictive gene sets using cDNA microarrays. *Mol. Pharmacol.*, **60**, 1189–1194.

Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. & Park, P.J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 13544–13549.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 6567–6572.

Tjin, E.P., Konijnenberg, D., Krebbers, G., Mallo, H., Drijfhout, J.W., Franken, K.L., van der Horst, C.M., Bos, J.D., Nieweg, O.E., Kroon, B.B., Haanen, J.B., Melief, C.J., Vyth-Dreese, F.A. & Luiten, R.M. (2011). T-cell immune function in tumor, skin, and peripheral blood of advanced stage melanoma patients: implications for immunotherapy. *Clin. Cancer Res.*, **17**, 5736–5747.

Tsoka, S. & Ouzounis, C.A. (2000). Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nat. Genet.*, **26**, 141–142.

Tsoka, S., Simon, D. & Ouzounis, C.A. (2004). Automated metabolic reconstruction for Methanococcus jannaschii. *Archaea*, **1**, 223–229.

Tsymbal A, C.P., Pechenizkiy M (2005). Diversity in search strategies for ensemble feature selection. *Information Fusion*, **6**, 83–98.

Turinsky, A.L., Razick, S., Turner, B., Donaldson, I.M. & Wodak, S.J. (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database (Oxford)*, **2010**, baq026.

Valeyev, N.V., Hundhausen, C., Umezawa, Y., Kotov, N.V., Williams, G., Clop, A., Ainali, C., Ouzounis, C., Tsoka, S. & Nestle, F.O. (2010). A systems model for immune cell interactions unravels the mechanism of inflammation in human skin. *PLoS Comput. Biol.*, **6**, e1001024.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Bernards, R. & Friend, S.H. (2003). Expression profiling predicts outcome in breast cancer. *Breast Cancer Res.*, **5**, 57–58.

Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D. & Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–245.

Vazquez, A., Dobrin, R., Sergi, D., Eckmann, J.P., Oltvai, Z.N. & Barabasi, A.L. (2004). The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 17940–17945.

Venter, J.C., Adams, M.D., Myers, E.W. & et al. (2001). The sequence of the human genome. *Science*, **291**, 1304–1351.

Vidal, M., Cusick, M.E. & Barabasi, A.L. (2011). Interactome networks and human disease. *Cell*, **144**, 986–998.

Vodovotz, Y., Csete, M., Bartels, J., Chang, S. & An, G. (2008). Translational systems biology of inflammation. *PLoS Comput. Biol.*, **4**, e1000014.

Volpe, E., Servant, N., Zollinger, R., Bogiatzi, S.I., Hupe, P., Barillot, E. & Soumelis, V. (2008). A critical function for transforming growth factor-beta, interleukin 23 and proinflammatory cytokines in driving and modulating human T(H)-17 responses. *Nat. Immunol.*, **9**, 650–657.

Wang, C.W. (2006). New ensemble machine learning method for classification and prediction on gene expression data. *Conf Proc IEEE Eng Med Biol Soc*, **1**, 3478–3481.

Wang, M., Chen, X. & Zhang, H. (2010). Maximal conditional chi-square importance in random forests. *Bioinformatics*, **26**, 831–837.

Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., Jatkoe, T., Berns, E.M., Atkins, D. & Foekens, J.A. (2005a). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.

Wang, Y., Makedon, F.S., Ford, J.C. & Pearlman, J. (2005b). HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics*, **21**, 1530–1537.

Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F. & Mewes, H.W. (2005c). Gene selection from microarray data for cancer classification–a machine learning approach. *Comput Biol Chem*, **29**, 37–46.

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

Wold, B. & Myers, R.M. (2008). Sequence census methods for functional genomics. *Nat. Methods*, **5**, 19–21.

Wolk, K., Kunz, S., Witte, E., Friedrich, M., Asadullah, K. & Sabat, R. (2004). IL-22 increases the innate immunity of tissues. *Immunity*, **21**, 241–254.

Wood, I.A., Visscher, P.M. & Mengersen, K.L. (2007). Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics*, **23**, 1363–1370.

Workman, C.T., Mak, H.C., McCuine, S., Tagne, J.B., Agarwal, M., Ozier, O., Begley, T.J., Samson, L.D. & Ideker, T. (2006). A systems approach to mapping DNA damage response pathways. *Science*, **312**, 1054–1059.

Xu, G. & Papageorgiou, L.G. (2009). A mixed integer optimisation model for data classification. *Comput. Ind. Eng.*, **56**.

Xu, M., Kao, M.C., Nunez-Iglesias, J., Nevins, J.R., West, M. & Zhou, X.J. (2008). An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. *BMC Genomics*, **9 Suppl 1**, S12.

Y., F.U., Yilmaz, O. & Turkay, M. (2008). Prediction of secondary structures of proteins using a two-stage method. *Computers amp; Chemical Engineering*, **32**, 78 – 88, ¡ce:title¿Process Systems Engineering: Contributions on the State-of-the-Art¡/ce:title¿ ¡ce:subtitle¿Selected extended Papers from ESCAPE '16/PSE 2006.¡/ce:subtitle¿.

Yan, Q. (2010). Translational bioinformatics and systems biology approaches for personalized medicine. *Methods Mol. Biol.*, **662**, 167–178.

Yang, F., Wang, H.Z., Mi, H., Lin, C.D. & Cai, W.W. (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC Bioinformatics*, **10 Suppl 1**, S22.

Yang, Y.e.a., P. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, **5**, 296–308.

Yao, Y., Richman, L., Morehouse, C., de los Reyes, M., Higgs, B.W., Boutrin, A., White, B., Coyle, A., Krueger, J., Kiener, P.A. & Jallal, B. (2008). Type I interferon: potential therapeutic target for psoriasis? *PLoS ONE*, **3**, e2737.

Yeh, I., Hanekamp, T., Tsoka, S., Karp, P.D. & Altman, R.B. (2004). Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome Res.*, **14**, 917–924.

Yeung, K.Y. & Bumgarner, R.E. (2003). Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome Biol.*, **4**, R83.

Yook, S.H., Oltvai, Z.N. & Barabasi, A.L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, **4**, 928–942.

Yu, G., Wang, L.G., Han, Y. & He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.

Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J. & Jarvis, E.D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, **20**, 3594–3603.

Zavaljevski, N., Stevens, F.J. & Reifman, J. (2002). Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics*, **18**, 689–696.

Zhang, X.J., Huang, W., Yang, S., Sun, L.D., Zhang, F.Y., Zhu, Q.X., Zhang, F.R., Zhang, C. & Du, e.a., W. H. (2009). Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21. *Nat. Genet.*, **41**, 205–210.

Zhao, S. & Bruce, W.B. (2003). Expression profiling using cDNA microarrays. *Methods Mol. Biol.*, **236**, 365–380.

Zheng, Y., Danilenko, D.M., Valdez, P., Kasman, I., Eastham-Anderson, J., Wu, J. & Ouyang, W. (2007). Interleukin-22, a T(H)17 cytokine, mediates IL-23-induced dermal inflammation and acanthosis. *Nature*, **445**, 648–651.

Zhou, N. & Wang, L. (2007). Effective selection of informative SNPs and classification on the HapMap genotype data. *BMC Bioinformatics*, **8**, 484.

Zhou, X., Krueger, J.G., Kao, M.C., Lee, E., Du, F., Menter, A., Wong, W.H. & Bowcock, A.M. (2003). Novel mechanisms of T-cell and dendritic cell activation revealed by profiling of psoriasis on the 63,100-element oligonucleotide array. *Physiol. Genomics*, **13**, 69–78.

Ziauddin, J. & Sabatini, D.M. (2001). Microarrays of cells expressing defined cDNAs. *Nature*, **411**, 107–110.

Zimdahl, H., Nyakatura, G., Brandt, P., Schulz, H. & Hummel, e.a., O. (2004). A SNP map of the rat genome generated from cDNA sequences. *Science*, **303**, 807.

Appendix A

Pseudocode for Random Forest Algorithm

Input: Training sets (x, y), Machine Learning Algorithm, Integer j (number of iterations)

Output: $g_n(x, N)$, tree classifier. Export Random Forest features: confusion matrix, variable importance, Out Of Bag (OOB) error, proximity matrix.

For $(i = 0; i < j; i++)$ do
Random Bootstrap Sampling with replacement of N observations
Bootstrap sample $D_n(N)$ is generated
Foreach sample calculate unpruned tree model. For M input variables, select $m << M$ from the original training sets and at each node the best split on the m is used to split the node. Generate a bagging classifier as $g^(m)_n(X, N_m, D_n)$ which is used as the voting classifier.
Foreach node do *(i)* the splitting using Gini Index (GI), *(ii)* sample m predictor variables and *(iii)* choose best partition using recursive partition method. **End Foreach**
End Foreach
End For

Appendix B

NetCFS algorithm is described here in more details by explaining its functions in java, WEKA API and Matlab as used from GENIE3.

Java Classes

Two dimensional arrays were used for a number of variables to store information regarding the networks.

1. public double datasetmatrix[][];: This array consists of a representation of the training dataset where rows represent observations and columns correspond to variables

2. public double rfmatrix[][], genematrix[][]; : $N \times N$ adjacent matrices, where N represents the number of genes in the training set. *Genematrix* keep the results from *CFSsubseteval* and *rfmatrix* represents the matrix obtained from Matlab

3. public double genenetwork[][]; : Similar as *Genematrix* but it stores the final structure of the gene expression network.

4. public void readarff () : Function that is used to open a .arff file (training set) to be parsed by weka object instances.

5. public void generate(); generate(String[] attributelist) : Using *CFS-subsetevaluation* from Weka API, a gene network is generated. A gene is selected as the predictor, class of the dataset, and next an object of type "attribute selection" from the WEKA API is created. This object is set to generate feature selection using the CFSsubset evaluation. The result of the feature selection is stored in an integer array holding the indexes of the genes selected. This array is then used to update the array genematrix.

6. public void generatewithrandomjungle(String rjdir, String tempre-

sultdir, String csvdir) : This method is used to calculate the weights (links) using random jungle. *rjdir*: random jungle executable directory,*tempresultdir*: directory storing the output and *csvdir*: directory to put the dataset in csv format. This method calls random jungle for each gene in the dataset.

7. public void exportnetworkcsv(double[][] resultmatrix, String csvdir): Function to export the gene expression network in csv format. Each row correspond to an edge and is described by three columns: the target node, the source node and the weight

8. public void generatehubs(double [][] tempgenematrix, String tempdir): Function to identify the hub genes. For directed networks the in-degree and out-degree edges of every node are calculated. Another version of this method have been created to find degree of undirected networks.

9. public void pccoexpressionnetwork(double cutoff, double negativecutoff): Function to generate a gene network using Pearson Correlation Coefficient (PCC). An adjacency matrix is created containing the correlation values between all genes. An undirected network is constructed keeping only the correlation value between two genes that is greater than a cut off being applied.

10. public void filterduplicates(): Function to remove duplicated edges by looking at every location in the adjacency matrix if item at $(i, j) == item(j, i)$ and their value is equal to 1. In this case the funtion removes the edge which has the smallest value calculated from GENIE3.

11. public void checksymetry(double[][] network) Function to count the number of edges which are symmetrical (edges which point to same nodes but differ in direction). The parameter is used to point to the network.

12. public void comparehubs (double [] list1, double[] list2, double cutoff, String tempdir) Comparison of hubs derived from two different networks. This method accepts two lists which contain the degree of each node. First the hubs of the list 1 are derived by selecting nodes which have number of neighbours greater than a cutoff value. Then the function exports the hubs to a CSV file and their respective number of degrees in the second list.

GENIE3

Here is given a brief description of the methods implemented in (Huynh-Thu

et al. [2010]) for GENIE3 and used by the algorithm presented in this thesis.

function VIM = genie3(exprmatrix,inputidx,treemethod,K,nbtrees)

This function is used to extract the weights for the construction of a network. Each setting has a different role: *exprmatrix:* A matrix where rows are samples and columns represent variables. *inputidx:* The list of genes which will be used to create the network. When the list is empty, all the genes are used. *treemethod:* type of the tree that was used *K:* number of features used in the splitting procedure *nbtrees:* number of trees created in the classification

function vi = genie3single(exprmatrix,outputidx,inputidx,treemethod,K,nbtrees)

Function to return a list of rankings for all genes with respect to the gene selected as a predictor. Parameters *exprmatrix,inputidx,treemethod,K and nbtrees* have the same function as in the function VIM, while *outputidx* is used as an index to the feature which will be used as the class/predicted value.

WEKA API

To do the classification and feature selection, WEKA API version 3.6 was used. This version contains a number of classes describing different objects and algorithms which can be accessed from a java application. A brief description of the items that were used is given below:

Object Instances

This object is used to keep the training datasets. Once the .arff file is read by the application it is automatically parsed and stored into an object of type instance. Each sample is stored as an object instance containing an array with the values for all variables.

Package classifiers

This package contains a number of classes which can be used as classifiers. These classes are subdivided by the classifier type which includes trees, functions and lazy methods. The classes name represents the algorithm they implement: Random Forest, Bagging, AdaBoostM1 (Boosting), LIBSVM (implementation used to generate support vector machines) and IBK (implementation of the K-nearest neighbour classifier).

Class Evaluation

Function to assess the evaluation of classifiers (for example 10-fold cross valida-

tion).

Attribute Selection

This object is used to do attribute selection on a given training dataset. All the different feature selection techniques are implemented in the package named attribute selection.